

Noise Correction on Subjective Datasets

Uthman Jinadu

Department of Computer Science,
Georgia State University
Atlanta GA 30303
ujinadu1@gsu.edu

Yi Ding

Department of Computer Science,
Georgia State University
Atlanta GA 30303
yiding@gsu.edu

Abstract

Incorporating every annotator’s perspective is crucial for unbiased data modeling. Annotator fatigue and changing opinions over time can distort dataset annotations. To combat this, we propose to learn a more accurate representation of diverse opinions by utilizing multitask learning in conjunction with loss-based label correction. We show that using our novel formulation, we can cleanly separate agreeing and disagreeing annotations. Furthermore, this method provides a controllable way to encourage or discourage disagreement. We demonstrate that this modification can improve prediction performance in a single or multi-annotator setting. Lastly, we show that this method remains robust to additional label noise that is applied to subjective data.

1 Introduction

The practice of enhancing label accuracy in datasets through the use of multiple annotations per sample is found in works such as [Snow et al. \(2008\)](#). This method capitalizes on the collective expertise of annotators to improve data quality. The reason is due to the assumption that individual annotators make mistakes and a collection of annotators can provide better-quality labels. Traditional methods for reconciling differences are to use techniques such as majority voting, averaging, or expert opinions ([Waseem and Hovy, 2016](#)). These methods are designed to arrive at a single ‘ground truth’ for training supervised learning models. However, in tasks of a subjective nature where a single ‘correct’ answer may not exist, these practices can compromise the diverse perspectives of individual annotators. These issues are found across many domains such as in medical ([Cheplygina and Plum, 2018](#)), social ([Ding et al., 2022](#)), and others ([Uma et al., 2021](#)).

To improve the fair representation of annotators’ opinions, datasets have been built to include indi-

vidual annotations. These datasets contain labels that are not aggregated through majority voting. This enables the capturing of opinions of annotators who would otherwise have been removed due to the aggregation function. Multitask learning is one way to model this data as demonstrated by [Mostafazadeh Davani et al. \(2022\)](#), where each task is to predict an annotator’s individual label. However, these methods do not account for possible mistakes that individual annotators might make during the annotation process. Datasets such as the GabHateCorpus by [Kennedy et al. \(2018\)](#) and GoEmotions by [Demszky et al. \(2020\)](#) can contain thousands of annotations by a single annotator. Mistakes such as misidentifying emotions due to annotation exhaustion would not be surprising. Furthermore, annotation of a large number of samples may occur over multiple days and introduce temporal distribution shifts of opinion.

We propose to address this problem by introducing loss-based label correction into a multitask learning setting. A fundamental property of loss-based label correction works by exploiting the network memorization effect by [Arpit et al. \(2017\)](#). They found that networks tended to learn simple patterns before learning more complex ones. This has been used extensively within the noisy learning community such as in [Arazo et al. \(2019\)](#). These methods first utilize the memorization effect to detect mislabeled instances. Then they correct these possible sources of mislabeling using techniques such as incorporation of a network’s own guess.

Motivated by this, we propose a novel formulation of multitask learning with label correction. However, a main challenge we try to overcome is that naive applications of these methods can erase the diverse perspectives of annotators. This is because the original technique makes use of a sample’s loss to determine whether it is correctly or incorrectly labeled. On a dataset with subjective labels, we find that higher loss samples are asso-

ciated with minority opinions which complicates the noisy sample detection process. We propose a novel method to address this by strengthening or weakening a model’s guess of the true label. We find that our method is robust to added noise while also maintaining diverse opinions. We show that this is useful when modeling subjective datasets with differing agreement properties. We highlight our contributions as follows:

- We present a novel formulation of multitask learning with loss-based noise correction.
- We demonstrate that we can separate agreeing and disagreeing annotations to detect noise and disagreement.
- We demonstrate that this modification can improve prediction performance in a single or multi-annotator setting. Additionally, this method remains robust to additional label noise that is applied to subjective data.
- We introduce a hyperparameter to control the degree of label correction due to variability in label properties for different annotation tasks. We show that this has a noticeable effect on performance.

2 Background

2.1 Multitask Learning for Disagreement

Disagreement among annotators, a common occurrence in the annotation process, signifies differing opinions or interpretations when labeling a specific task. Addressing annotator disagreement is a fundamental part of enhancing machine learning model accuracy. Traditionally, reliability has been used as a measure of dataset quality, however, this measure emphasizes having a single ground truth. This is problematic for capturing diverse opinions. To address these problems, researchers have studied various underlying sources of such disagreements, focusing on two primary types: random variation and systematic disagreement (Krippendorff, 2011; Dumitrache, 2015; Aroyo and Welty, 2013).

In many real-life scenarios, obtaining the true label for training—considered the gold standard or objective ground truth—is impractical, too costly, or tedious. We often rely on labels from multiple sources, which may introduce noise and significant variation. This variation underlines the challenge of applying supervised learning algorithms without a definitive gold standard, especially when label

sources vary in reliability and accuracy (Raykar et al., 2010).

Random variation encompasses the inherent unpredictability in human judgments, which can introduce randomness into machine learning algorithms, impacting their outcomes (Fernandes et al., 2023). This type of disagreement might stem from individual biases or the subjective nature of a task such as due to demographic properties (Ding et al., 2022).

On the other hand, systematic disagreement is more nuanced, arising from consistent inaccuracies or biases in the machine learning models themselves. Systematic disagreement can also arise from multiple sources or due to multiple reasons. As demonstrated by Krippendorff, this can result from systematic factors that skew the models, leading to persistent errors in interpretation (Krippendorff, 2008).

Understanding both types of disagreement is crucial for developing robust machine learning models, as it improves the reliability of annotations. Classifier performance and resilience are influenced by the level of annotators’ agreement, which is important when selecting training and test data (Leonardelli et al., 2021). Ongoing research continues to enhance methods for identifying and modeling disagreements, thereby improving the quality and representativeness of annotated data.

2.2 Annotator Noise

Noise correction refers to the systematic process of identifying and rectifying errors or inconsistencies within a dataset, often termed error correction. Its primary objective is to enhance both the quality and reliability of the data by either eliminating or minimizing the influence of noise (Zhan et al., 2019).

This noise may include random errors, artifacts, or inaccuracies that can significantly distort the underlying patterns and relationships within the data. By employing various correction techniques tailored to the specific type and source of the noise, researchers can obtain a cleaner, more accurate representation of the information.

Traditional categorical cross-entropy loss is not well-suited for this task as it tends to fit the noise rather than the underlying data distribution (Zhang et al., 2016). The hard bootstrapping loss introduced by Reed et al. (2015) offers a solution by augmenting the standard cross-entropy loss with a perceptual term. This addition helps adjust the training objective, making it more robust to label

noise.

Deep neural networks demonstrate a two-phase learning process when trained on datasets with noisy labels. In the "early learning" phase, they tend to pick up correct and general patterns. As training continues, they start to memorize incorrectly labeled examples. This behavior reflects their initial ability to capture generalizable data patterns before gradually fitting to noise, which challenges their generalization capability on new data. This property has been used by many works to improve representation learning under noisy conditions (Arazo et al., 2019; Liu et al., 2020; Li et al., 2020; Nishi et al., 2021).

Effective noise correction is essential in data preprocessing, especially in machine learning, where data quality significantly impacts model performance and interpretability. This ongoing challenge involves carefully eliminating genuine noise without losing critical variations or introducing new biases (Arazo et al., 2019).

2.3 Uncertainty Quantification

Machine learning models' uncertainty is particularly useful in subjective areas, helping to determine when human oversight is needed, such as in content moderation, where uncertain predictions trigger human review to maintain content standards. (Ghandeharioun et al., 2019) (Chandrasekharan et al., 2019)

Estimating uncertainty in machine learning is crucial but challenging. The simplest method uses Softmax probabilities (Hendrycks and Gimpel, 2017), but it can yield overconfident predictions for novel inputs, unlike the training data.

Gal's Monte Carlo dropout technique, highlighted in Gal and Ghahramani (2016), improves uncertainty estimation by applying dropout not only during training but also during testing. This method calculates output variances to offer a more accurate assessment of prediction uncertainty than single-label probabilities.

Klas and Vollmer (2018) argues that relying solely on single label probabilities overlooks many elements influencing predictive uncertainty. Instead, employing a range of predicted annotations provides a more refined and thorough assessment of uncertainty in machine learning models.

3 Methods

We are given dataset denoted as D , comprising text samples X , and an annotation matrix $Y \in \mathcal{R}^{2 \times A}$. Here, X represents individual text entries, A signifies the number of annotators, and Y is a matrix where each cell y_i stores one hot value of a label class. Due to the non-uniform distribution of annotator input across text instances, missing values in Y are prevalent. Our goal is to deduce a consensus label for each text instance.

3.1 Baseline

A standard way for optimizing a model for multi-class classification is to utilize the cross entropy loss function. For an input sample x , the cross entropy loss is given by:

$$L_{CE} = - \sum_{m=1}^M y_m \cdot \log(p_m(x, \theta)) \quad (1)$$

where m represents the number of classes and the y_i represents the truth for the i th class. θ signifies some model.

3.2 Multi-task learning to account for different opinions

Individual annotators' opinions can be incorporated into learning the model via multi-task learning. Suppose there are A annotators, we can create A different task heads (fully connected layers) to predict each of the individual annotations. This changes the cross-entropy loss to a summation of the individual outputs for each task head:

$$L_{MT} = - \frac{1}{A} \sum_{a=1}^A \sum_{m=1}^M y_{a,m} \cdot \log(p_{a,m}(x, \theta)). \quad (2)$$

Here, the ground truth values for each sample form a $A \times M$ matrix \mathbf{y} . The model will also output a prediction distribution for each annotator. Note that MT loss reduces to the CE loss when the number of annotators is 1.

3.3 Noise Correction via Loss Modeling

Arazo et al. (2019) found that a sample's loss can be used to separate correctly and incorrectly annotated samples. That is, when learning with imperfect data, lower losses corresponded to the correct samples, and higher losses correlated with incorrect annotations. This leads to a two-step algorithm. First, the loss distribution was modeled using a two-component mixture model to generate a weight w .

The weight indicates how likely a particular sample, based on its loss, is likely to be mislabeled. The intuition is that a model’s guess should be integrated for possibly mislabeled data. This gives us a loss function which scaled by this weight:

$$L_{LC} = -((1 - w) \sum_{m=1}^M y_i \cdot \log(p_m(x, \theta)) + w \sum_{m=1}^M z \cdot \log(p_m(x, \theta)))$$

The new label is re-parameterized as a weighted sum of a network’s guesses with its ground truth. Here, $z = p_m(x, \theta)$, and represents a network’s prediction based on a single round of evaluation and is detached during a training cycle of the network.

3.4 Multi-task learning with loss-based noise correction

We can now combine equations L_{MT} and L_{LC} from above to create a multi-task label correction loss:

$$L_{MLC} = \frac{1}{A} \sum_a -((1 - w) \sum_{m=1}^M y_{a,m} \cdot \log(p_{a,m}(x, \theta)) + w \sum_{m=1}^M z_{a,m} \cdot \log(p_{a,m}(x, \theta)))$$

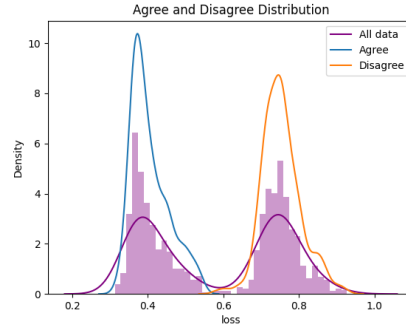
$$L_{MLC} = -\frac{1}{A} \sum_{a=1}^A \sum_{m=1}^M (1 - w) y_{a,m} \cdot \log(p_{a,m}(x, \theta)) - \frac{1}{A} \sum_{a=1}^A \sum_{m=1}^M w z_{a,m} \cdot \log(p_{a,m}(x, \theta))$$

$$L_{MLC} = (1 - w) * (L_{MT}) + (\psi * w * L_G)$$

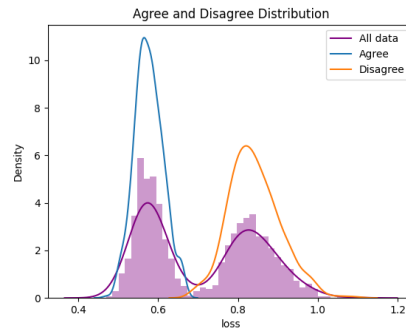
Additionally, we propose to introduce a subjectivity parameter ψ parameter to tune the impact of a network’s self-guess loss term $\psi \cdot L_G$. Intuitively, as shown in Figure 1, higher values would push more disagreeing labels to the majority class. This can be tuned as a way to balance agreement with disagreement and noise based on dataset properties.

3.5 Manifold Mixup

Integrating the manifold mixup technique into our baseline and multitask loss correction scenarios, as delineated by Verma et al. (2018), involved strategically selecting a random layer from the BERT model. This choice facilitated the application of mixup at the embedding level, addressing the challenges of working with textual data which inherently limits direct input interpolations. Manifold mixup alters the representations linked to layers preceding and succeeding the mixing procedures.



(a) GabHateCorpus loss distribution for a single annotator.



(b) GoEmotions loss distribution for a single annotator.

Figure 1: Modeling the loss distribution for annotators in multitask learning for a given epoch of training. Arpit et al. (2017) demonstrated that networks tended to learn correct samples first before memorizing the incorrect samples. In our problem, the semantics of this changes to agreeing versus disagreeing samples. As can be seen, there is a clear separation of samples based on loss. For the dataset depicted in Figure 1a, the average proportion of samples in the "Agree" category is 54.57%, while the "Disagree" category averages 45.43%. Within the "Agree" distribution, 55.43% of the samples conform to the baseline majority. In contrast, within the "Disagree" distribution, 48.94% of the samples align with the majority view. For the dataset illustrated in Figure 1b, the "Agree" category comprises an average of 51.69% of the samples, whereas the "Disagree" category averages 48.31%. Within the "Agree" distribution of the GoEmotions sample, 52.84% of the samples align with the baseline majority. Conversely, in the "Disagree" distribution, 48.65% represent the majority view. Additionally, this demonstrates that our method can preserve diverse opinions. This also conforms with the previous understanding that networks tend to be over-confident with predictions since the lower losses are more aligned than higher losses with the majority case.

4 Experimental Setup

We evaluate each of the four different optimization methods above in these experiments.

- **Baseline** corresponds to a network with a sin-

gle output head optimized to the majority label using L_{CE} .

- **Baseline + Loss Correction** injects our loss correction technique onto the majority label optimization method (L_{LC}). The manifold mixup method was adopted.
- **Multitask** involves adding A fully-connected layers to the last layer of the network. Each output is trying to predict the annotations of a unique annotator. Since the datasets we are using are missing annotators, these loss terms are zero-ed out and not back-propagated. This corresponds to the (L_{MT}) loss.
- **Multitask + Loss Correction** is our proposed method for integrating loss correction (L_{MLC}) into a multitask setting. Just like the baseline loss correction scenario, we applied the manifold mixup to this as well.

4.1 Datasets

In our experiments, we made use of GoEmotions and GabHateCorpus datasets to annotate our subjective tasks.

GoEmotions (Demszky et al., 2020) is a large-scale dataset consisting of 58K Reddit comments annotated with 27 different emotion categories, including neutral labels. The dataset is annotated by 82 crowdsourced workers and label noise is minimized by removing labels that have fewer than one annotator. The dataset is designed for multi-class and multi-label emotion classification, however, 83% of examples have a single label. The dataset is pre-split into a training set with 43,410 entries, a validation set containing 5,426 entries, and a testing set comprised of 5,427 entries. We evaluate the six basic Ekman emotion labels in this work.

GabHateCorpus (GHC): The Gab Hate Corpus (GHC) hate speech dataset developed by Kennedy et al. (2018), and consists of 27,665 text entries collected from Gab.com. This public dataset was annotated by 18 specialized annotators, with each entry evaluated by at least 3 different experts. This rigorous process ensures a high level of accuracy and consistency within the GHC. This property is unique due to its emphasis on label quality when given label diversity.

4.2 Implementation Details

The classification models were built using the HuggingFace transformers library (version 4.38) (Wolf

et al., 2019). Our experimental setup closely resembled that of Mostafazadeh Davani et al. (2022). For the GabHateCorpus experiments, we trained the models for five epochs using a learning rate of $1e-7$. As for the GoEmotions dataset, we utilized a learning rate of $5e-6$ and trained the models for five epochs. We used this to obtain the baseline and multitask for each of the datasets and then we introduced the label correction similar to how Arazo et al. (2019) was implemented. We used the pre-trained *Bert base-uncased* model as the underlying architecture. As we do not have access to the original multi-task code some results are different than Mostafazadeh Davani et al. (2022). Optimization was conducted using Stochastic Gradient Descent (SGD) with a momentum of 0.9, weight decay of 0.01, and a batch size of 64.

Additionally, we integrate some components that are necessary to train for loss correction. A warm-up period using an entropy-based penalty on the confidence term has been found useful to prevent noise overfitting during warm-up periods (Pereyra et al., 2017). We designated a warm-up period of 2 epochs for the GoEmotions and GabHateCorpus datasets. This was necessary to stabilize the training, mitigate poor initialization, and make it start with a smaller learning rate and gradually increase it. We utilized regularization following Tanaka et al. (2018) and Arazo et al. (2019) aiming to deter the allocation of all samples to a singular class. We did not implement the full Arazo et al. (2019) to minimize possible confounding effects of auxiliary techniques. Unless otherwise specified, ψ is set to 0.5.

4.3 Noise Injection

The datasets used in the evaluation are for binary classification. Therefore, we simply flip the labels by sampling for a subset of data. In this paper, we used 20% noise rate for the noise experiments. For samples where the labels are from multiple annotators, e.g. 4 votes for true and 1 vote for false, we swap the labels for the samples where we are injecting noise. This is similar to how labels are injected in popular noisy labels literature. Not all annotators annotated every sample. We ignored these annotators when injecting noise and thus these did not have an impact on the loss.

Text	Annotator Rating	Majority	Agree With Majority?	Multitask
Yeah, OK then, get your nurse to give you some more of those nice tablets that make you sleep.	Not Hate	Not Hate	Yes	Not Hate(100%)
This creature was predicated his actions were predicted etc etc etc now the media are outraged as predicted.... this is proof that the level of evil we are dealing with is not a person but a large and well financed group	Not Hate	Not Hate	Yes	Not Hate(100%)
Farcebook? Should have tanked a long time ago.	Not Hate	Not Hate	Yes	Not Hate(67.67%)
IDK, my favorite was "Icelandic Women are Whores" Happy New Year and Hail Victory	Not Hate	Not Hate	Yes	Not Hate(100%)
Christian cake baker who won SCOTUS case is now being harassed by demented Satanist homosexuals, backed by the state of Colorado.	Not Hate	Hate	No	Hate(50%)

Table 1: Samples from the GabHateCorpus Dataset which shows the text, the annotation on the text given a specific annotator, agreement with the majority based on the Agree and Disagree distribution in Figure 1, the baseline majority with the percentage of Hate in the annotation. The samples show instances where the baseline differs removes minority opinions that multitask captures.

Text	Emotion	Annotator Rating	Majority	Agree With Majority?	Multitask
But also: fuck the daily mail.	Anger	Anger	Anger	Yes	Anger(100%)
You blew it. They played you like a fiddle	Disgust	Not Disgust	Not Disgust	Yes	Not Disgust(80%)
The sin was refusing to impregnate his brother	Disgust	Not Disgust	Disgust	No	Disgust(66.67%)
Ok now that was epic!	Fear	Not Fear	Not Fear	Yes	Not Fear(100%)
I can see why! I would totally be offended too!	Sadness	Sadness	Not Sadness	No	Not Sadness(66.67%)
I've been abused enough by the Knicks today.	Sadness	Sadness	Sadness	Yes	Sadness(100%)
I just can't believe she had the nerve to call.	Surprise	Surprise	Surprise	Yes	Surprise(100%)

Table 2: Samples from the GoEmotion Dataset which shows the text, the annotation on the text by a specific annotator compared to the baseline majority with the percentage of the emotion, and if it agrees with the majority based on the Agree and Disagree distribution in Figure 1. The samples show instances where the baseline distribution is different than the multitask setting.

5 Results

Due to the nature of this problem, it is important to consider these results within the context of data collection processes. GoEmotions annotators were crowdworkers who spoke English from India. GHC annotation was a two-step process that involved expert opinions. The annotators of GHC were trained and the annotation process was much more controlled. Furthermore, the annotator population was recruited from a smaller sample and consisted of undergraduate students. These processes would cause GHC to be less variable, due to likely similar backgrounds of annotators. On the other hand, using untrained crowdworkers from a foreign country would likely lead to more variability.

With this in mind, we first discuss the comparison of performance metrics in evaluated methods. We then examine the effect of injecting additional noise. Lastly, we examine the effect of adjusting the subjectivity parameter and its effect on results.

5.1 Comparison of techniques

We present the GabHateCorpus results in Table 3 and the results of GoEmotions in Table 4. All metrics reported are compared to the majority label on the ground truth. As can be seen in the GHC results, our method demonstrates improvements across all metrics when compared against the baselines. Additionally, we see that baseline methods based on the majority also perform worse than baseline + label correction. These results indicate that regardless of the number of annotators, it may be beneficial to incorporate label correction.

For GoEmotions results, we see that for Multitask+LC, all six cases exceed the multitask results. This indicates that the loss correction is having some effect. Most notably, there is a significant improvement in performance between the majority and annotator (multitask) conditions. We hypothesize that this is due to two factors. One plausible explanation for this phenomenon may be the inconsistency in the number of annotators who labeled each text instance in the dataset. This pattern was

Method	No Noise			20% Noise		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	46.09 ± 2.9	49.96 ± 5.7	47.95 ± 5.4	49.87 ± 5.6	49.78 ± 5.6	41.67 ± 4.7
Baseline + Label Correction	50.11 ± 3.2	51.32 ± 3.3	49.34 ± 5.6	50.50 ± 3.2	50.12 ± 3.2	45.97 ± 5.2
Multitask	49.56 ± 5.6	49.86 ± 5.0	45.57 ± 2.9	47.70 ± 3.0	47.95 ± 3.1	45.62 ± 2.9
Multitask + Label Correction	51.54 ± 3.3	51.71 ± 2.8	50.3 ± 3.2	52.32 ± 3.3	51.12 ± 2.5	51.55 ± 2.3

Table 3: Performance comparison on the GabHateCorpus dataset when compared to majority based ground-truth. Our proposed method of incorporating label correction into multitask learning had the highest score. Our method is also robust to label noise under conditions of label subjectivity. This experiment is run with 5 different seeds and we report the averages.

also noted by Mostafazadeh Davani et al. (2022). Such variance could introduce an additional layer of complexity or uncertainty into the model, affecting its ability to efficiently generalize from the data. This opens up avenues for future research to understand how the number of annotators, and the possible divergence in their annotations, influences the effectiveness of loss correction methods in multi-task learning setups. While we observed an improvement in the multitask scenario, we believe the application of manifold mixup is also contributory to this, our multitask results exceed the results presented in the previous work. Secondly, additional investigation is warranted for the effect of the initial model which further confounds the results.

5.2 Classification of Subjective Labels in the Presence of Noise

We compare the effect of noise for each classification task on GHC as well as GoEmotions. In this paper, we conduct experiments with a 20% noise rate injected based on the process discussed in 4.3. As can be seen in Table 3 our method is robust to noise injection even when accounting for multiple annotators. It is also interesting to note that both the majority-based methods for aggregating labels (Baseline Majority and Baseline + Label Correction), showed a large drop in performance. This highlights the importance of training multiple annotators as a way to mitigate the effect of noise.

We see that for the results of GoEmotions in Table 5, there is a significant drop in the majority label annotations. However, the multitask cases all showcase a smaller drop in performance than the majority label techniques. This highlights the need to account for multiple annotator opinions.

5.3 Effect of Subjectivity Parameter

We present the results exploring the subjectivity parameter ψ in Table 6. A higher ψ will lead to more conforming predictions, while a lower ψ will lead to less conforming predictions. Our investigation into the optimal setting for the subjectivity parameter (ψ) across GHC and GoEmotions datasets demonstrates that a ψ value of 0.5 consistently achieves the highest performance (relative to the majority vote) for our experimental setup. This result highlights the importance of a balanced approach to noise correction across diverse datasets, showing that a ψ value of 0.5 effectively manages label disagreements and boosts model accuracy. In addition to the majority-based patterns we see in Table 6 we present annotator-level results in Table 7. As can be seen, there is a direct correlation between the annotator-level labeling variance and the ψ parameter. This appears to be more visible in the dataset GHC as it is more balanced than GoEmotions. To calculate variance across all annotations, we follow the same formula as previous work:

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

The inherent noise in the dataset suggests that some correction of annotation mistakes is necessary, but it will compromise prediction diversity. The results reported are averaged over two runs.

While this value may change depending on hyperparameters, dataset, architecture, and other properties; it highlights the need to take a balanced approach when adjusting for noise correction on the dataset with subjective annotations. Setting the value too high may lead the network to make incorrect guesses and may lead to increased variance and inaccuracies. Setting it to an ‘optimal’ value may cause the dataset to predict the majority cases more, however, this may not be desirable in all sit-

No Noise				
Emotion	Baseline	Baseline + LC	Multitask	Multitask + LC
Anger	42.21 ± 3.7	43.35 ± 3.6	56.31 ± 3.6	66.68 ± 4.6
Disgust	32.4 ± 6.1	34.18 ± 3.1	58.16 ± 2.4	64.65 ± 0.5
Fear	44.45 ± 6.0	49.09 ± 5.7	58.96 ± 5.5	69.80 ± 6.1
Joy	52.17 ± 5.3	54.63 ± 5.0	60.91 ± 5.2	66.97 ± 4.4
Sadness	50.29 ± 5.7	51.84 ± 3.8	62.16 ± 5.8	67.86 ± 3.9
Surprise	44.32 ± 5.1	46.86 ± 6.1	61.63 ± 5.0	68.93 ± 3.5

Table 4: Performance comparison on the GoEmotions dataset. Binary classification accuracy is reported for each emotion. For each emotion considered, we demonstrate improved performance when ψ is fixed to 0.5.

20% Noise				
Emotion	Baseline	Baseline + LC	Multitask	Multitask + LC
Anger	46.58 ± 5.9	49.93 ± 3.0	44.28 ± 5.0	48.26 ± 3.0
Disgust	44.3 ± 3.3	48.94 ± 4.9	43.89 ± 1.1	46.27 ± 4.9
Fear	44.38 ± 3.6	49.08 ± 3.5	44.11 ± 4.8	47.05 ± 3.1
Joy	46.61 ± 4.9	50.35 ± 6.1	43.62 ± 5.2	48.83 ± 3.6
Sadness	45.51 ± 3.6	49.47 ± 4.9	43.58 ± 3.8	47.77 ± 1.9
Surprise	44.53 ± 1.2	49.24 ± 2.9	43.83 ± 3.6	49.01 ± 3.4

Table 5: Performance comparison on the GoEmotions dataset when injected with 20% label noise. Our method shows robustness to noise injection as well. The observed performance reduction in multitask and multitask+LC models, compared to the baseline and baseline+LC, upon noise injection might be attributed to the specific method of integrating noise, the number of annotators of the GoEmotion dataset, and other inherent characteristics of the GoEmotion dataset. This suggests exploring alternative noise simulation techniques might mitigate the impact on model performance, particularly in datasets with complex attributes.

uations and thus our ψ parameter provides a way to control for this property.

5.4 Loss Separation

We showcase a clear bimodal distribution to illustrate the challenge of separating noise and opinion in Figure 1. We see that for a specific annotator, there are peaks for agreeing and disagreeing samples. Note that these are semantically different than the noisy label’s definition of correct and incorrect. In this paper, we refer to incorrect annotations as a label that the annotator would not typically ascribe to a sample due to the dataset properties. We assume that incorrect answers are more likely to fall in the distribution of disagreeing labels. We calculate the proportion of labels within each mode to see how many of them agree with the majority annotation and how many disagree. Our proposed technique attempts to control the level of noise correction to account for these added complications.

5.5 Model Uncertainty

In our empirical evaluations, we observed an increase in the agreement between model predictions and majority labels upon applying the loss correction strategy to multitask scenarios. The provided graphs in Figure 2 illustrate the effect of the loss correction method on the variance of predictions for two multitask models across different values of ψ .

For the GHC Multitask model depicted on the left of Figure 2, there is a visible trend where the variance decreases to its lowest point at $\psi = 0.5$. This represents an 18% increase in the agreement between the model’s predictions and the majority labels. The GoEmotion Multitask model shown on the right shows a corresponding improvement of approximately 14% improvement in prediction consistency. The other points on the graph for GoEmotion vary more significantly, with some ψ values leading to higher variances, which suggests greater fluctuation in agreement levels at those points.

Both graphs highlight the optimal setting of ψ to 0.5 for achieving the most reliable predictions, as indicated by the minimum variance values at this point, suggesting the effectiveness of the loss correction strategy in enhancing model reliability. This process quantifies the consistency of model predictions where a lower variance suggests a higher agreement with majority. This highlights the technique’s effectiveness in boosting the models’ reliability and applicability, especially in scenarios characterized by subjective judgment and high uncertainty.

6 Discussion

Noisy-label tolerant techniques make the assumption that there is a single ground truth. Many of these techniques, e.g. [Arazo et al. \(2019\)](#), utilize a label’s loss to determine whether a sample is correctly labeled or not. That is higher loss is generally incorrectly labeled. However, when opinion is involved, we cannot naively do this. As we show in our results on Figure 1, the semantics of a samples’ loss changes to majority or minority opinions. That is, higher loss is associated with minority opinions while lower loss is associated with majority ones.

At a per-annotator level, there are still mistakes being made. Thus we need to find a way to utilize these loss properties when learning representations. To do so, we propose a novel loss function (Section 3.4) which models the predictions for each

Label type	No Noise			20% Noise		
	$\psi = 1$	$\psi = 0.5$	$\psi = 0.25$	$\psi = 1$	$\psi = 0.5$	$\psi = 0.25$
GHC	47.86	50.3	48.13	45.8	51.55	48.71
Anger	55.05	66.68	56.96	45.40	48.26	45.49
Disgust	56.74	64.65	55.5	44.64	46.27	44.82
Fear	57.55	69.80	57.55	44.80	47.05	43.67
Joy	55.57	66.97	57.83	44.45	48.83	44.53
Sadness	53.42	67.86	53.41	44.68	47.77	44.73
Surprise	55.39	68.93	60.77	45.36	49.01	43.74

Table 6: Comparison of accuracy for GHC and GoEmotions for different parameters of ψ . Adjusting ψ has a clear effect on GHC and GoEmotions. Both GHC and GoEmotions appear to perform the best in general for $\psi = 0.5$. We believe the application of the manifold mixup to the training has some effects on this. This indicates a difference in the label patterns for subjectivity and noise.

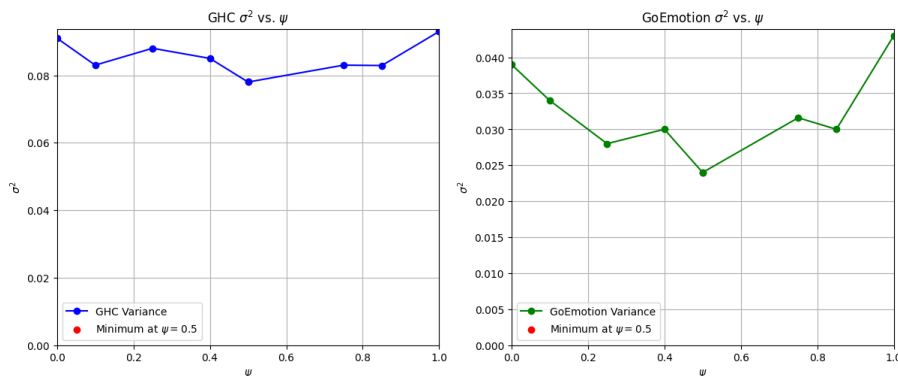


Figure 2: The variance in model predictions against majority labels for the GHC Multitask and GoEmotion Multitask models. Notably, at a ψ value of 0.5, variance reaches its minimum, indicating the highest agreement among annotators. This point reflects the optimal balance in the loss correction strategy, enhancing model alignment with consensus sentiment.

Dataset	Metric	$\psi = 0.25$	$\psi = 0.5$	$\psi = 1$
GHC	Anno. Acc.	59.48	63.82	58.67
GHC	Pred. Var.	14.38	13.59	13.01
GoEmo	Anno. Acc.	49.92	50.11	49.85
GoEmo	Pred. Var.	17.74	17.53	17.46

Table 7: Annotator-level accuracy and variance when adjusting for ψ reported as percentages. Lower ψ encourages diversity of annotation while higher ψ encourages agreement. As can be seen, prediction variance is highest when the ψ value is lowest. The annotator-level accuracy is also the highest when the $\psi=0.5$ and reflects the same patterns seen in majority vote results shown in Figure 2. These results demonstrate that we can tune for opinion and noise properties in a controllable way.

annotator which factorizes out nicely. We then introduce an additional parameter to control the strength of this noise correction process to balance noise with disagreeing opinions. This methodology is particularly relevant for subjective tasks like online hate speech detection on platforms such as Reddit and Twitter, striving for models that effec-

tively reduce noise and enhanced opinion modeling. Our experimental outcomes underscore the significance of adopting a balanced strategy for noise correction across varied datasets (Table 7). It also shows cases that annotator-level mistakes exist and can be corrected for (Tables 3, 4, 5). Accounting for these factors leads to an improvement in performance metrics at both annotator and aggregate levels while maintaining prediction diversity.

7 Conclusion

In this study, we developed a method for loss-based label correction for a multitask setting. Specifically, this is applied to subjective datasets from multiple annotators. We show that our approach preserves disagreeing labels and that the performance increases under some conditions. We also demonstrated that it is robust to added noise. Lastly, we demonstrate that adjusting the degree of loss correction is important and has a large impact on performance. This is beneficial for future work in collecting subjective labels at scale.

8 Limitations

Although ψ significantly affects performance, particularly at 0.5, identifying the optimal value remains complex without a clear selection strategy. The parameter reduced variance and improved performance, as evidenced in the GoEmotions dataset, but interpreting these gains is difficult due to the GoEmotions dataset imbalance. Yet, our results still show an improvement over previous studies (Mostafazadeh Davani et al., 2022). This underscores the need for further research to develop a systematic approach for determining ψ , crucial for the method’s adaptability and precision across different datasets. Additionally, the typical way to analyze the noise patterns of less subjective datasets is to annotate a subset of the test set at scale. This gives insight into the noise profile and can enable the creation of a label noise transition matrix. To fully evaluate the noise patterns of subjective annotations is much more costly. Annotation subjectivity can vary through time, and with demographic properties. These properties make building a noise transition matrix almost impossible due to the number of factors that can influence the underlying label distribution. Thus, understanding noise for subjective annotation tasks remains an exciting challenge we leave for future work.

9 Ethics Statement

In data annotation, capturing the full spectrum of annotator perspectives is important for achieving equitable outcomes. Challenges such as annotator fatigue and variability in judgment over time can obscure the true diversity of opinions within large datasets. To address these issues, our method introduces a novel combination of multi-task learning and loss-based label correction. This method effectively separates matching and mismatching annotations while improving prediction accuracy for both individual and multi-annotator settings. Similar to the use of multi-annotator models in affect detection (Alm, 2011), our method models diverse opinions for a deeper insight into subjective data. Additionally, our technique adapts to the intrinsic label noise in subjective annotations, enhancing robustness similar to methods in sarcasm detection that consider annotator-specific thresholds (Rakov and Rosenberg, 2013). This refined approach captures the complexity of human judgment, promoting fairer data interpretation in digital annotation.

10 Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Research was also supported by a gift award from The Home Depot. We thank the reviewers for their constructive criticism and valuable advice, which have improved the quality of this final paper.

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.
- E. Arazo, D. Ortego, P. Albert, N. E., and K. McGuinness. 2019. [Unsupervised label noise modeling and loss correction](#). *ArXiv*.
- Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).
- Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 233–242. JMLR.org.
- Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelie, 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.
- 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 ExpertLabel Synthesis*, pages 105–111. Springer.
- D. Demszky, J. Ko, A. Cowen, G. Nemade, and S. Ravi. 2020. [Goemotions: A dataset of fine-grained emotions](#). *ArXiv preprint ArXiv:2005.00547*.
- Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

- Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 701–710. Springer.
- Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668.
- 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.
- 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.
- Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*.
- Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv*. July, 18.
- Michael Klas 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.
- Klaus Krippendorff. 2008. Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measure*, 2(4):323–338.
- Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication methods and measures*, 5(2):93–112.
- E. Leonardelli, S. Menini, A. P. Aproso, M. Guerini, and S. Tonelli. 2021. [Agreeing to disagree: Annotating of-fensive language datasets with annotators’ disagreement](#). *ArXiv*.
- J. Li, R. Socher, and S. C. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *ArXiv*, abs/2002.07394.
- Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. [Early-learning regularization prevents memorization of noisy labels](#). In *Advances in Neural Information Processing Systems*.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. 2021. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031.
- Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.
- Rachel Rakov and Andrew Rosenberg. 2013. "sure, i did the right thing": a system for sarcasm detection in speech. In *Interspeech*, pages 842–846.
- Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermsillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of machine learning research*, 11(4).
- Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint*.
- 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.
- Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5552–5560.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, and Yoshua Bengio. 2018. Manifold mixup: Better representations by interpolating hidden states. *arXiv preprint arXiv:1806.05236*.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. 2019. [Learning from multi-annotator data: A noise-aware classification framework](#). *ACM Trans. Inf. Syst.*, 37(2).
- Chiyan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.