When Disagreement Meets Noise: Noise Robust Annotator Embeddings for Subjective NLP

Anonymous ACL submission

Abstract

Subjective NLP tasks such as sentiment analysis and hate speech classification often involve inherent annotator disagreement, reflecting diverse perspectives shaped by annotators' lived experiences. Although conventional approaches resolve disagreement through majority voting or aggregation, these methods risk erasing valuable nuances and minority viewpoints. Recent embedding-based/multitask models have advanced the modeling of annotator-specific judgments, yet their robustness to annotation noise remains underexplored. In this work, we systematically investigate how state-of-the-art disagreement learning models perform in the presence of noisy labels and observe a significant performance degradation under such conditions. To address this, we propose Noise Robust Annotator Embedding (NRA-Embed), which integrates Robust InfoNCE (RINCE) contrastive loss to enhance models' robustness under noisy annotation conditions. Moreover, we benchmark existing approaches across three axes: label noise type (symmetric vs. rogue annotators), task structure (binary vs. multiclass), and annotator coverage (many vs. few labels per example). Through extensive experiments, we show that NRA-Embed effectively models subjective variation while remaining resilient to noise, achieving competitive or superior performance compared to prior methods.

1 Introduction

Collecting multiple annotator judgments is standard in NLP to improve label reliability (Snow et al., 2008; Nowak and Rüger, 2010). Disagreement in ground truth annotations arise frequently and are typically resolved through majority voting, averaging (Sabou et al., 2014), or expert adjudication (Waseem and Hovy, 2016) to create a single ground truth for supervised training. However, for subjective tasks, where no "correct" label exists (Alm, 2011), forcing a single annotation can obscure valuable nuances in annotators' diverse perspectives (Cheplygina and Pluim, 2018).

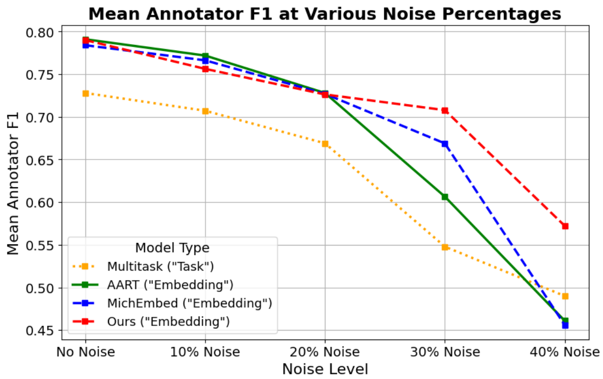

Figure 1: Comparison of Mean Annotator F1 scores for different multi-annotator modeling approaches on the MDA(Leonardelli et al., 2021) Dataset under increasing noise levels. We highlight that our embedding-based method demonstrates greater robustness, maintaining higher performance compared to other approaches as noise level increases.

Annotators' backgrounds and experiences shape subjective annotations in tasks like political stance detection (Luo et al., 2020), sentiment analysis (Díaz et al., 2018), and hate speech identification (Patton et al., 2019). Feminist and anti-racist activists differ from crowd workers on hate speech (Waseem and Hovy, 2016), and political affiliations affect neutrality (Luo et al., 2020). Majority voting risks suppressing minority views, causing disparities (Prabhakaran et al., 2021). These differences in opinions, or disagreements, are important to capture in datasets to build safe and fair models.

Many datasets have been built to understand annotation disagreement, such as the Multi-Domain Agreement (MDA) dataset (Leonardelli et al., 2021). Many techniques have also been proposed to better capture this annotator disagreement, which can arise from errors, ambiguous items, or subjective opinions, often tied to lived experiences (Uma et al., 2021; Reidsma and Carletta, 2008;

Uma et al., 2022). For example, Davani et al. (2022a) proposed a multitask approach for doing so. Jinadu and Ding (2024) extended this to tolerate errors while accounting for subjective properties. Embedding-based approaches such as AART (Mokhberian et al., 2023), and work by Deng et al. (2023) have pushed the state of the art in terms of accuracy on disagreement benchmark datasets. However, few works have systematically examined how models behave under the presence of both annotator disagreements as well as inaccuracies.

To understand these properties, we systematically examined how state-of-the-art disagreement learning methods behave in the presence of label noise. We found that there is a sharp performance drop in multi-annotator models under noise, as seen in Figure 1. To address these shortcomings, we propose the **Noise Robust Annotator Embedding method (NRA-Embed)**, which incorporates Robust InfoNCE (RINCE) (Chuang et al., 2022) to optimize disagreement learning models. In addition, we conducted extensive experiments to evaluate models across noise types, classification settings, and annotator conditions. We found that our NRA-Embed approach is effective for learning under disagreement and noise, demonstrating performance that is at least on par with, and often surpasses, existing state-of-the-art methods.

2 Background

2.1 Inherent Annotator Disagreement

Annotator disagreement constitutes a recognized challenge in Natural Language Processing. Conventional methodologies for addressing this issue include label aggregation through averaging techniques (Pavlick and Callison-Burch, 2016), implementing majority voting systems (Sabou et al., 2014), or selectively utilizing data subsets characterized by high inter-annotator agreement levels (Jiang and de Marneffe, 2019). However, assuming a single "correct" label ignores genuine subjectivity, multiple valid interpretations exist (Plank, 2022; Passonneau et al., 2012; Nie et al., 2020; Jiang and Marneffe, 2022).

Evidence from several studies demonstrates that genuine variability in human annotations can stem from subjective interpretations or the existence of multiple acceptable responses (Passonneau et al., 2012; Nie et al., 2020; Jiang and Marneffe, 2022). For example, in tasks such as toxic language detection, perceptions of toxicity vary significantly among individuals (Waseem, 2016; Al Kuwatly et al., 2020). Annotators' identities and personal beliefs substantially shape their judgments about the toxicity of content (Sap et al., 2021). Thus, differences among annotators should not merely be treated as annotation "noise" (Pavlick and Kwiatkowski, 2019). Recent work has begun utilizing these diverse annotations to personalize models more effectively for different users (Plepi et al., 2022).

2.2 Modeling Annotator Disagreement

Multiple methods address annotator disagreement. The classical Dawid-Skene model (Dawid and Skene, 1979) uses an expectation-maximization algorithm to estimate annotator reliability and the underlying true labels jointly from noisy labels. More recently, Zhang and de Marneffe (2021) introduced Artificial Annotators to simulate uncertainty, and Zhou et al. (2021) applied MC Dropout, Deep Ensembles, Re-Calibration, and Distribution Distillation to capture judgment variability. Meissner et al. (2021) modeled full label distributions for the Natural Language Inference (NLI) task. Zhang et al. (2021) handle mixed single-label, multi-label, and unlabeled data. Gordon et al. (2022) propose "jury learning" via a DCN (Wang et al., 2021) that integrates text and annotator IDs, while Davani et al. (2022a) add annotator-specific layers on top of a shared representation, and Kocoń et al. (2021) learn per-annotator embeddings.

Zhang et al. (2021) explores annotator disagreement within a broader context involving a combination of single-label, multi-label, and unlabeled examples. Meanwhile, (Gordon et al., 2022) propose "jury learning," a method that individually models annotators using a Deep and Cross Network (DCN) (Wang et al., 2021). Their approach integrates the textual input and annotator identifiers along with predicted annotator responses from DCN for improved classification outcomes.

2.3 Annotator Noise

Noise correction seeks to identify and resolve errors or inconsistencies ("noise") in datasets, such as random label flips, artifacts, and annotation mistakes, to improve data quality and model robustness (Zhan et al., 2019). Standard cross-entropy losses tend to fit noise rather than the true underlying distribution (Zhang et al., 2016), whereas "hard" bootstrapping augments the loss with a prediction-based term to resist label noise (Reed et al., 2014).

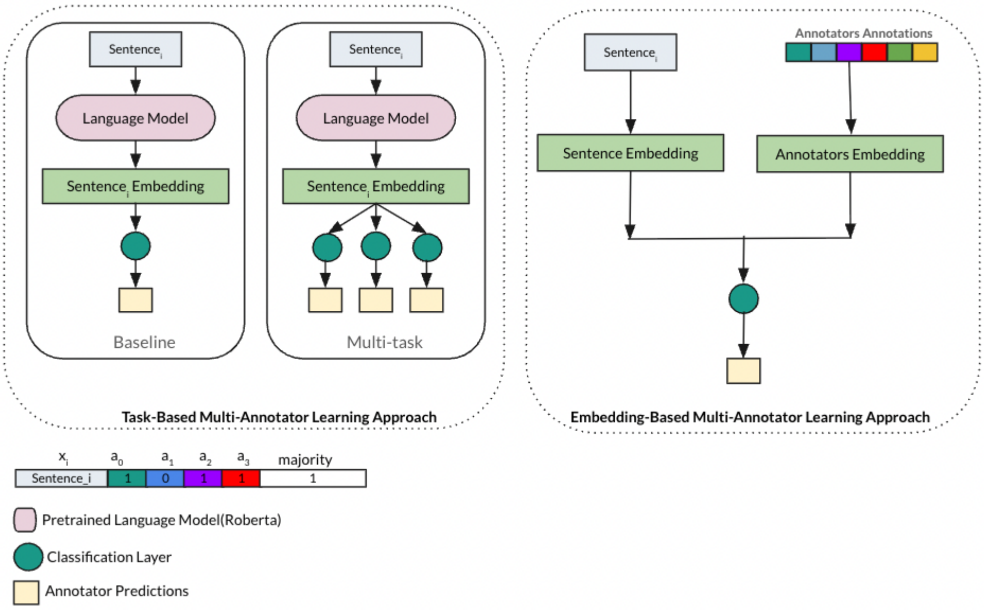

Figure 2: Architectural approaches for modeling multi-annotator learning. Two main methods exist: First, the task-based approach (left) adds specific prediction layers to capture individual annotators' perspectives in a subjective dataset. The embedding-based approach (right) incorporates annotator information by embedding their annotations directly into a latent representation early in the network.

Empirically, deep nets first learn broad, generalizable patterns before eventually memorizing noisy labels (Arazo et al., 2019; Liu et al., 2020; Li et al., 2020; Nishi et al., 2021), a phenomenon that many methods exploit to cleanse noisy data. Effective noise correction thus balances removal of misleading errors against preservation of genuine signal, avoiding new biases in downstream models (Arazo et al., 2019; Jinadu and Ding, 2024).

3 Methods

We first set up the general multi-annotator learning problem. We then discuss two high-level network architecture approaches for modeling multi-annotator datasets. At 3.2, we discuss the task-based approach, as depicted in the left section of figure 2, where a separate shallow sub-network predicts individual annotator responses. We then discuss the embedding-based approach at 3.3, as depicted in the right section of the figure 2, where an embedding is learned for each annotator to represent their annotating preference. We then propose a framework for optimizing the embedding-based approaches under conditions of label noise.

3.1 Problem Definition

We consider an annotated dataset $D = \{(x_i, a_j, y_{ij})\}$, which consists of triplets formed from input text items $X = \{x_i\}_{i=1}^N$, annotators $A = \{a_j\}_{j=1}^m$, and annotations $Y = \{1, ..., Q\}$. A pair of (i, j) can appear at most once in the dataset D, which means label y_{ij} is assigned to text item x_i by the annotator a_j. Y contains numerous missing values in most annotated datasets since each annotator labels only a subset of the instances. The problem is modeled as a classification task where a classifier is trained to predict the label to be assigned to the text item x_i. All methods explored in this study utilize pre-trained transformer-based language models for text encoding, specifically using RoBERTa (Liu et al., 2019). For a given input text x_i, we obtain its text representation by extracting the [CLS] token embedding from the final layer of the language model, denoted as $e(x_i)$.

3.2 Task-based Multi-Annotator Learning

The most frequent approach, called ***single-task***, aims to predict the aggregate label to be assigned to the text item x_i through majority voting or averaging over annotators' labels $\{y_{ij}\}_{j=1}^M$. It is typically implemented by passing the text representation of a pre-trained BERT-base language model $e(x_i)$ through a fully connected layer. This layer performs a linear transformation followed by a Softmax activation to produce the probability distribution over the majority of labels.

Multi-task approaches are seen in several previous works (Fornaciari et al., 2021; Davani et al., 2022a; Jinadu and Ding, 2024), and basically are a generalization of single-task approach. They train a separate, fully connected layer for each annotator to learn the annotator-specific labeling behavior. They leverage shared pre-trained BERT-base language model (Liu et al., 2019) (encoding layers) to produce a unified text representation $e(x_i)$ for all annotators. However, these shared encoding layers are updated jointly using the outputs from all annotator-specific tasks. The training objective for each annotator is defined independently using a cross-entropy loss, applied only to the labels that the annotator has provided for each instance x_i.

3.3 Embedding-based Multi-Annotator Learning

Another method is to embed annotators in a latent space and integrate this information early in the model architecture. In these approaches, a learnable matrix encodes the representations of the annotators. During training, annotators which provided the rating can be retrieved from the embedding matrix and inserted into the network. For example, given a text instance x_i and an annotator embedding, we compute the annotator-aware embedding as:

$$g(x_i, a_j) = e(x_i) \oplus f(a_j),$$

where $e(x_i)$ is the text embedding, $f(a_j)$ is the corresponding annotator embedding, and $\oplus$ is the fusion operation that can arise from a linear layer, attention or something more complex. A model then processes this fused representation to determine the optimum prediction. In our paper we treat the fusion as a simple addition.

A few methods make use of this. For example, the approach proposed by Mokhberian et al. (2023) adds the annotator embeddings directly into the text representations without any weighting. We refer to this method as Annotator Aware Representations for Texts(**AART**) in this paper. Another method is Deng et al. (2023), which additionally incorporates annotation embeddings along with weighting. We refer to this method as **MichEmbed** in this paper.

3.4 Noise Robust Annotator Embedding (NRA-Embed)

We found that embedding-based approaches performed better than task-based approaches. However, it is unclear how to make these methods more noise-robust while capturing subjective opinions. These challenges are illustrated in Figure 1; in noisy environments, conventional contrastive losses such as InfoNCE (Oord et al., 2018; Chen et al., 2020) often fail to learn embeddings that accurately reflect annotators' true opinions. Because inconsistent or noisy annotations can distort the learning signals and hinder the model's ability to form coherent representations. Therefore, we need a contrastive loss robust to annotation noise—tunable to emphasize confident, informative annotation signals while downweighting uncertain or potentially noisy ones. Motivated by this, we propose to use Robust InfoNCE (RINCE) (Chuang et al., 2022).

RINCE builds on the insight that contrastive learning with noisy representations can be interpreted as a binary classification with noisy labels over pairwise views—assigning a label of 1 if the views co-occur (joint distribution) and -1 if sampled independently (product of marginals)(Chuang et al., 2022). This interpretation aligns well with our setting, where each view corresponds to an annotator's label on a given input; we treat annotator pairs as positive (label 1) if they agree on the label of a text instance and negative (label -1) if they disagree. Ghosh et al. (2015) demonstrate that *symmetric* loss functions offer robustness to label noise in binary classification tasks. RINCE introduces a *symmetric adaptation of contrastive learning* that satisfies the symmetry condition in binary classification and, thus, guarantees robustness against noisy representations. Specifically, a symmetric contrastive learning objective should have the following form(Chuang et al., 2022):

$$\mathcal{L}(s) = \underbrace{\ell(s^+; 1)}_{\text{Positive Pair}} + \lambda \sum_{i=1}^{K} \underbrace{\ell(s_i^-; -1)}_{K \text{ Negative Pairs}} \quad (1)$$

where the first term is the loss of the positive pair, and the second term is the sum of losses of K negative pairs. $\lambda > 0$ is a density weighting term controlling the ratio between positive (class 1) and negative (class -1) pairs.

Based on the idea of robust symmetric classification loss, the Robust InfoNCE (RINCE) loss is defined as(Chuang et al., 2022):

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q}(s) = \frac{e^{q \cdot s^+}}{q} + \frac{\left(\lambda \cdot \left(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}\right)\right)^q}{q} \quad (2)$$

where s^+ is the score for a positive (agreement) pair and s_i^- are scores for negative (disagreement) pairs. A tunable parameter $q \in (0, 1]$ is introduced to interpolate between the robustness of RINCE and the expressive power of InfoNCE. When $q = 1$, RINCE becomes a contrastive loss that fully satisfies the symmetry property in Equation (1) and offers strong resistance to annotation noise.

To jointly learn task performance and annotator embeddings, we pass combined embeddings $g(x_i, a_j)$ through a classification layer to predict the annotator's label for each instance. we optimize the following comminatory objective function:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \sum_j \|f(a_j)\|_2^2 + \lambda_2 \sum_{j,j'} \mathcal{L}_{\text{RINCE}}(j, j') \tag{3}$$

The first term, $\mathcal{L}_{\text{CE}}$, is a standard cross-entropy loss used to predict the label assigned by annotator a_j to input x_i based on the combined embedding $g(x_i, a_j)$. The second term applies an ℓ_2 regularization penalty on the annotator embeddings $f(a_j)$, encouraging smoother and more generalizable representations. The third term incorporates the RINCE contrastive loss between pairs of annotators a_j and $a_{j'}$ who have labeled the same text instance. Annotator pairs who agree on a label are treated as positives and pulled together in the embedding space, while those who disagree are pushed apart—encouraging consistency while maintaining robustness to noisy annotations.

4 Experimental Setup

4.1 Datasets

We use the following datasets in our evaluation.

- **The Multi-Domain Agreement Dataset (MDA)** This dataset comprises 9,814 English tweets drawn from three topical domains (the Black Lives Matter movement, the 2020 U.S. election, and the COVID-19 pandemic), each independently annotated for offensiveness by five crowdworkers via Amazon Mechanical Turk (Leonardelli et al., 2021).

- **Sentiment Analysis Dataset (SNT)** The dataset, introduced by Díaz et al. (2018), is a sentiment classification resource aimed at addressing age-related biases in sentiment models, leveraging text from older adults' blog posts containing age-related terms such as "old" and "young".

- **HS-Brexit Dataset (HSB)**
The HS-Brexit Dataset (HSB), introduced by Akhtar et al. (2021) is a multi-perspective abusive language detection dataset focused on Brexit-related tweets in English. It captures diverse viewpoints, especially from victimized groups like immigrants, with annotations for hate speech, aggressiveness, offensiveness, and stereotypes. Annotations were performed by varied demographic groups, including migrants, and a polarization index (P-index) was used to measure differing perspectives, creating separate gold standards per group. The dataset enabled training of perspective-aware models, including BERT-based classifiers, to better detect abusive language by considering annotator subjectivity. It serves as a benchmark for studying abusive language detection and sociodemographic biases in polarizing contexts like Brexit.

4.2 Baseline Models

We compare against the following baseline methods:

- **Multitask:** We follow the approach proposed by (Davani et al., 2022b) which involves one fully-connected layer for each annotator with a shared RoBERTA model.

- **AART:** We evaluate the approach introduced by (Mokhberian et al., 2023) which utilizes an embedding for each annotator as well as a contrastive loss objective and a single fully-connected classification layer built off of a RoBERTA backbone in our evaluations. Embedding-based approach.

- **MichEmbed:** We follow the approach by (Deng et al., 2023) which utilizes annotator embeddings as well as weighted annotation embeddings and a single fully-connected classification layer built off of RoBERTA in our experiments. Embedding-based approach.

4.3 Noise Injection

For our evaluations on binary-label datasets we evaluate noise by introducing label flips ("noise") into a random subset of examples. Specifically, we injected noise rates of 20% and 40%. For each selected instance, regardless of how many annotators originally voted for "true" versus "false" (e.g., 4 votes true, 1 vote false), we simply swapped its label. This approach mirrors the standard random-flip procedures commonly used in the noisy labels literature. Any annotators who did not contribute to a given sample were excluded from the noise injection process and thus did not affect the training loss.

For the multi-class dataset SNT, we add symmetric noise for each annotator of a label. Each instance labeled by an annotator had a 20% or 40% chance of being flipped. In this case a "flipped" label would result in one of the other 4 classes with equal likelihood.

4.4 Implementation Details

We implemented the classification models using HuggingFace transformers library(version 4.39) (Wolf et al., 2020). Our experimental setup for the annotation embedding approach for subjective classification closely resembled that of Mokhberian et al. (2023). For all the datasets experiments, we trained the models for ten (10) epochs. We used this to train our baseline and the other models, and then introduced our noise correction method. We used the pretrained Roberta-base (Liu et al., 2019) model as the underlying architecture. Optimization was conducted using the AdamW optimizer with a learning rate of 1e-5 and a weight decay of 0.01. A linear decay scheduler with zero warm-up steps was then applied.

4.5 Evaluation Metrics

Mean-Annotator F1 Score

This study is driven by the need to preserve minority annotator perspectives that are often lost when labels are naively aggregated. To that end, we evaluate the model's performance for each annotator a_j on each test item x_i, comparing the true labels y_{ij} against the model's predictions. We then summarize these per-annotator results via the **Mean-Annotator F1**, defined as the average macro-F1 score across all J annotators:

$$\text{Mean-Annotator F1} = \frac{1}{J}\sum_{j=1}^{J} \text{F1}\big(a_j\big),$$

where $\text{F1}\big(a_j\big)$ is the macro-F1 score computed for annotator a_j over all test items x_i.

Accuracy Score

The accuracy metric for our annotator-aware representation model quantifies the overall fraction of correct label predictions across every item–annotator pair (x_i, a_j) in the test set. Concretely, if $\hat{y}_{ij}$ denotes the model's predicted label for (x_i, a_j) and y_{ij} is the true label provided by annotator a_j, then accuracy is given by

$$\text{Accuracy} = \frac{\big|\{(i,j) \mid \hat{y}_{ij} = y_{ij}\}\big|}{\big|\{(i,j)\}\big|},$$

where the numerator counts all correctly predicted pairs and the denominator is the total number of evaluated pairs. This global correctness measure complements our annotator-wise F1 scores by showing, at a glance, how often the model's annotator-specific representations produce the right label across both prolific and sparse contributors.

5 Results

5.1 Main Results

We present the main results in Table 1 as mean-annotator f1 scores. We we see that our method performs the best on several metrics. Most notably it outperforms existing techniques in no presence of noise as well. Embedding approaches consistently outperform the multitask model, demonstrating their superior ability to capture individual annotator behaviors. We observe a large variation on performance in SNT.

5.2 Annotation Embedding Approach + Noise Robustness Enhancements

We compare our NRA-Embed Approach against the annotation embedding approach baseline to measure gains in embedding stability under synthetic annotation noise. Table 1 presents mean Annotator-Aware F1 scores on three benchmarks, MDA, SNT, and HSB, at no noise, 20% noise, and 40% noise. Our Noise-Robust Annotation Embedding Approach consistently improves over the baseline, demonstrating enhanced robustness to annotation errors.

5.3 Impact of Parameter q on Noise-Robustness in Annotation Embedding

Higher values of q in our Noise-Robust Annotation Embedding approach improve performance under noisy conditions. As q increases, the approach places greater emphasis on confident (easy) positive pairs while reducing the influence of noisy, ambiguous positives. This aligns with the intuition that contrastive learning objectives should be more selective in identifying trustworthy signals under

Dataset	Noise Level	Majority Vote	Multitask	MichEmbed	AART	Ours
MDA	No Noise	0.582(accuracy)	0.728	0.784	<u>0.788</u>	**0.790**
	20% Noise	–	0.669	0.727	<u>0.728</u>	**0.751**
	40% Noise	–	<u>0.490</u>	0.456	0.451	**0.572**
SNT	No Noise	0.303(accuracy)	0.287	**0.524**	0.452	<u>0.493</u>
	20% Noise	–	0.253	**0.440**	<u>0.421</u>	0.410
	40% Noise	–	0.217	**0.355**	<u>0.335</u>	0.300
HSB	No Noise	0.772(accuracy)	0.929	<u>0.933</u>	0.931	**0.936**
	20% Noise	–	<u>0.875</u>	0.872	0.833	**0.877**
	40% Noise	–	**0.674**	0.594	0.580	<u>0.663</u>

Table 1: Annotator-level F1 scores across three datasets (MDA, SNT, HSB) under varying levels of synthetic label noise. Best results are in **bold** and second best are <u>underlined</u>. Our method performs best or second best in most conditions, especially under high-noise conditions (The q-value chosen varies depending on what provided the best results).

higher noise levels. However, excessively high q values may overlook legitimate harder cases, indicating a trade-off between robustness and representational richness, particularly in low-noise scenarios. In practical settings with inconsistent crowdworker annotations, higher q values (e.g., $q = 0.75$ and $q = 1.0$) have proven reliably effective. This trend is illustrated in Table 2, where increasing q enhances the model's confidence and robustness.

5.4 Impact of Renegade Annotators on Model Performance

We analyze model robustness in realistic scenarios involving renegade annotators, individuals who intentionally provide malicious or random annotations. To do this, we randomly choose 10% of annotators to have very high noise, that is 70% of their annotations are perturbed. Experiments compare our proposed Noise-Robust Annotation Embedding method against the Task-Based approach. Results demonstrate that our method falls short in this being noise robust with few instances of high noise. Future work should explore how to handle these sorts of annotators. Detailed performance metrics are provided in Table 3.

6 Discussion

Our results demonstrate several key trends that hold consistently across datasets and noise configurations, offering both theoretical and practical insight into designing models for subjective classification under noisy annotation.

RINCE consistently improves model robustness across noise scenarios. Across most tested noise levels, our approach led to a notable increase in mean annotator F1 and reduced degradation under high-noise conditions. This supports our hypothesis that subjective NLP tasks require not only modeling of annotator identity but also a mechanism to counteract annotation noise.

Annotator embedding models outperform multitask learning. Our results show that models such as MichEmbed, AART and, our approach NRA-Embed, which learn annotator embeddings to modulate shared representations, outperform multitask approaches with separate prediction heads per annotator. We hypothesize that this advantage arises from parameter sharing and regularization effects—embedding-based models can exploit commonalities across annotators while still personalizing behavior, whereas multitask heads may overfit when annotation coverage is sparse or imbalanced. Additionally, embedding approaches inherently support more efficient transfer across annotators and can generalize better when annotators have limited individual data.

Likert Scale based datasets degrade contrastive loss performance An interesting finding is the results of SNT which is based on a Likert Scale classification. We find that previously strong approaches like AART and our approach, degrade in performance. We hypothesize that this is due to their objective being dependent on contrastive loss. For example, a contrastive loss would treat

	No Noise			20% Noise			40% Noise		
Rince_q	q = 0.5	q = 0.75	q = 1.0	q= 0.5	q = 0.75	q = 1.0	q= 0.5	q = 0.75	q = 1.0
MDA	0.7900	0.7850	0.7825	0.725	0.7417	0.7512	0.6405	0.6495	0.6572
SNT	0.4872	0.4907	0.4934	0.3991	0.4034	0.4103	0.2944	0.2947	0.2969
HSB	0.9337	0.9321	0.9359	0.8516	0.8588	0.8566	0.5998	0.6517	0.5882

Table 2: Effect of the parameter q on the robustness of the Noise-Robust Annotation Embedding method across different noise levels. We can see a trend that higher q-values tend to improve performance in higher noise scenarios.

Model	MDA	SNT	HSB
Task-Based	0.666	0.229	**0.862**
NRA-Embed	0.710	0.425	0.854
AART	**0.730**	0.434	0.853
MichEmbed	0.711	**0.438**	0.858

Table 3: Comparison of model robustness to renegade annotators (malicious/random annotation behavior). Bolded values highlight the best-performing approach across datasets.

DS	#A	#E/#A	#S	#L
MDA	819	60	44k	2
SNT	1481	41	60.4k	5
HSB	6	952	5.7k	2

Table 4: Dataset Statistics. #A is the number of antators, #E/#A is the average number of examples per annotator, #L is the number of possible labels, and #S is the total number of samples in the dataset. We obtain values from (Deng et al., 2023).

labels "Strongly Agree" and "Moderately Agree" as a negative pair in the same way it would consider "Strongly Agree" and "Strongly Disagree" as negative pairs. This is likely what led to a drop in the contrastive loss performance. On the other hand, an approach like MichEmbed which relies on a combination of Annotator + Annotation Embeddings performs strongly. Future works should look into modifying contrastive loss to be more class-sensitive such as in the case of Likert-based classification.

Multitask models degrade in performance with sparse annotators. The multitask model performed the worst with the SNT dataset. This is likely due to how many annotators there are compared to how many samples they annotated on average, which is very few, creating sparse annotators (see Table 4). On the other hand, the multitask model performed very well on HSB which had a much smaller amount of annotators who each labeled many samples.

These findings reinforce the need to view subjective learning as a two-fold challenge: embracing disagreement while resisting noise. Annotator-aware models alone are not sufficient if they assume all disagreement is meaningful; conversely, noise-robust objectives without subjectivity modeling may conflate diverse opinions with error. Our work shows that integrating both perspectives yields the most reliable performance, and that simple but principled interventions—like swapping InfoNCE for RINCE—can offer significant gains in real-world annotation environments.

7 Conclusion

In this work, explored the distinction of label noise and subjective disagreement in subjective learning tasks. Most prior works only consider one or the other; however, these factors are intertwined. Disagreement is core to many human-centered activities and should be accounted for when building datasets. We address this issue by separating label disagreement and label noise through our NRA-Embed approach. Our benchmarking of existing multi-annotator models provides a strong baseline for developing advanced models that can tolerate unique noise patterns. Our results suggests that embedding based approaches are the superior methodology for training in multi-annotator cases. Furthermore, we recommend that raw labels should be released, however noisy, so that issues with label noise can be directly addressed by model.

8 Limitations

One primary limitation of our approach is that synthetic noise cannot be a true replacement for real-world noise in our evaluations. In future works, it may be worthwhile to explore various types of noise-injection that more accurately reflect real-world noise. Another limitation is that per-annotator modeling may be a computationally expensive task, especially in datasets with large amounts of annotators, further research should be explored on grouping annotators or on other mechanisms to reduce this.

9 Ethics Statement

In data annotation, capturing the full spectrum of annotator perspectives is crucial for producing fair and representative models. However, factors like annotator fatigue and shifting judgments over time can conceal the true range of opinions present in large datasets.

To address this, we propose drawing on insights from the entire annotator pool—including those who contribute less frequently—rather than focusing solely on the most active contributors. Incorporating these “sparser” judgments broadens the diversity of viewpoints the model sees, yielding predictions that are both more robust and more nuanced.

That said, this inclusive approach carries its own risks. A small, coordinated subgroup of annotators might exert undue influence, and any biases embedded within our large language model infrastructure could further distort individual annotations. Even so, we argue that the benefits of embracing a wider array of voices—enhancing both inclusivity and resilience in AI systems—far outweigh these potential drawbacks.

10 Acknowledgements

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

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.

Eric Arazo, Diego Ortego, Paul Albert, Noel O’Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Veronika Cheplygina and Josien PW 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: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 105–111. Springer.

Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. 2022. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16670–16681.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022a. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022b. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

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.

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.

Aritra Ghosh, Naresh Manwani, and PS Sastry. 2015. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213.

Uthman Jinadu and Yi Ding. 2024. Noise correction on subjective datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5385–5395, Bangkok, Thailand. Association for Computational Linguistics.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE international conference on data mining (ICDM)*, pages 1168–1173. IEEE.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of nli models. *arXiv preprint arXiv:2106.03020*.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

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.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rebecca J Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46:219–252.

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.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. *arXiv preprint arXiv:2210.14531*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

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

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

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

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5:818451.

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.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pages 1785–1797.

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.

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. 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.

Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. 2019. Learning from multi-annotator data: A noise-aware classification framework. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–28.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. *arXiv preprint arXiv:2109.04408*.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2021. Distributed nli: Learning to predict human opinion distributions for language reasoning. *arXiv preprint arXiv:2104.08676*.