

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

Kevin Jacob¹ Parsa Ghazvinian² Uthman Jinadu²
Nicole Tianjiao Yang³ Yi Ding³

¹Georgia Institute of Technology ²Georgia State University

³University of Tennessee, Knoxville

kjacob7@gatech.edu pghazvinian1@student.gsu.edu

ujinadu1@gsu.edu {tyang23, yding}@utk.edu

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 and multi-task 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 multi-annotator learning models perform in the presence of noisy labels, and find that embedding-based approaches are more robust than aggregation and multitask baselines on datasets with many sparse annotators. Furthermore, we propose a new methodology, Noise-Robust Annotator Embedding (NRA-Embed), which integrates Robust InfoNCE (RINCE) contrastive loss to enhance models' robustness under noisy annotation conditions. Through extensive experiments, we demonstrate that NRA-Embed effectively captures subjective disagreement while maintaining robustness to noise, achieving performance that is comparable to state-of-the-art in clean settings while providing numerically greater robustness under noise.

1 Introduction

Collecting multiple annotations per instance is a standard practice in NLP to improve label reliability and mitigate individual annotator errors (Snow et al., 2008; Nowak and Rüger, 2010). In conventional supervised learning, disagreement among annotators is typically resolved through majority voting, averaging (Sabou et al., 2014), or expert adjudication (Waseem and Hovy, 2016) to produce a single ground-truth label. However, for *subjective NLP tasks*, such as sentiment analysis, hate

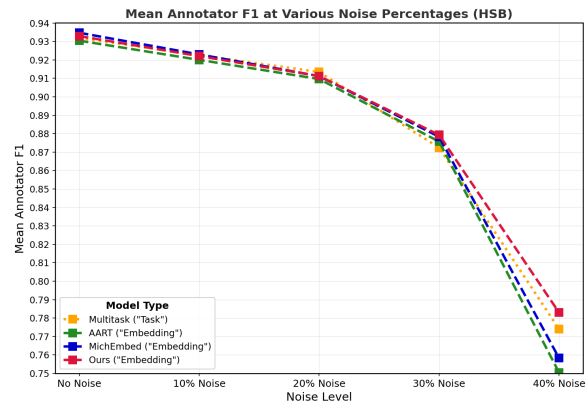


Figure 1: Comparison of Mean Annotator F1 scores for different multi-annotator modeling approaches on the HSB (Akhtar et al., 2021) dataset under increasing noise levels. We highlight that our embedding-based method often demonstrates greater robustness, maintaining higher performance compared to other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

speech detection, and political stance classification, a single objectively correct label often does not exist (Alm, 2011). Enforcing consensus in these settings can obscure meaningful variation in human judgments and suppress socially situated perspectives (Cheplygina and Pluim, 2018). At the same time, real-world annotation pipelines are inevitably affected by *annotation noise*, arising from misunderstanding, inattention, or malicious behavior. Despite this, the joint effect of subjective disagreement and annotation noise remains insufficiently understood.

Annotator disagreement in subjective tasks is not random but is often systematically shaped by annotators' demographic backgrounds, social identities, and political beliefs. For instance, feminist and anti-racism activists have been shown to annotate hate speech differently from crowd workers (Waseem and Hovy, 2016), while political affiliations influence labeling behavior in stance detection (Luo et al., 2020). Similar effects have been observed in sentiment analysis (Díaz et al., 2018)

and toxic language annotation (Patton et al., 2019). Collapsing such diverse judgments into a single label via majority voting risks marginalizing minority viewpoints and introducing representational harms (Prabhakaran et al., 2021). Motivated by these concerns, a growing body of work has sought to *model* annotator disagreement rather than eliminate it. Benchmark datasets such as Multi-Domain Agreement (MDA) (Leonardelli et al., 2021) enable systematic study, while probabilistic reliability modeling, large-scale disagreement learning, and distributional supervision (Reidsma and Carletta, 2008; Uma et al., 2022, 2021) in prior works provide principled alternatives to label aggregation.

Building on these foundations, recent methods have moved beyond label aggregation to explicitly model annotator perspectives. These include multitask architectures with per-annotator classification heads (Jinadu et al., 2023; Mostafazadeh Davani et al., 2022; Jinadu and Ding, 2024) and embedding-based approaches that represent annotators as learned vectors (Mokhberian et al., 2024; Deng et al., 2023). Embedding methods using contrastive learning have achieved state-of-the-art performance on disagreement benchmarks under clean supervision (Mokhberian et al., 2024).

However, contrastive objectives like InfoNCE fundamentally rely on correct positive/negative pair construction (Oord et al., 2018) and when labels are corrupted, the contrastive signal becomes unreliable. In this work, we conduct the first systematic study of how state-of-the-art disagreement learning methods behave under controlled annotation noise. We show empirically in Figure 1 that methods which account for multi-annotator perspectives degrade in performance with increasing annotation noise. These findings indicate that existing embedding-based disagreement models are not inherently robust to noisy supervision.

To address this limitation, while keeping the benefits of contrastive loss, we argue that robustness must be incorporated directly into the contrastive learning objective itself. We therefore propose the **Noise-Robust Annotator Embedding method (NRA-Embed)**, which is an annotator embedding-based model that replaces the conventional InfoNCE loss (Oord et al., 2018) with *Robust InfoNCE (RINCE)* (Chuang et al., 2022). RINCE introduces a principled robustness mechanism that reduces sensitivity to corrupted positive and negative pairs through a tunable robustness parameter. By integrating RINCE into the annotator em-

bedding framework, NRA-Embed preserves the benefits of contrastive annotator modeling while improving robustness to annotation noise.

We evaluate NRA-Embed across multiple subjective classification tasks, noise types, and annotator conditions. Our contributions are:

- We conduct the first systematic evaluation of multi-annotator models under controlled noise, revealing significant vulnerabilities in existing approaches.
- We propose NRA-Embed, a noise-robust variant of contrastive embedding based models that matches state-of-the-art in clean settings while providing numerical advantages under noise.
- We benchmark across three binary datasets and one multiclass dataset and two noise types (symmetric and instance-dependent), establishing best practices for noise-robust subjective learning.

2 Related Works

2.1 Annotator Disagreement

Annotator disagreement is a recognized challenge in NLP. Traditional approaches treat this variability as noise through majority voting or label aggregation (Pavlick and Callison-Burch, 2016; Sabou et al., 2014). However, many tasks admit multiple valid interpretations where no single correct label exists (Jinadu et al., 2023; Plank, 2022; Passonneau et al., 2012; Nie et al., 2020; Jiang and Marneffe, 2022; Jinadu and Ding, 2024). In subjective settings like toxicity detection, judgments differ due to annotators’ beliefs and identities, reflecting genuine human diversity rather than errors (Waseem, 2016; Al Kuwatly et al., 2020; Sap et al., 2022; Pavlick and Kwiatkowski, 2019). Recent work leverages this disagreement for personalization (Plepi et al., 2022).

2.2 Modeling Annotator Disagreement

Methods for modeling disagreement have evolved from treating it as noise to embracing it as signal. Early approaches include Dawid–Skene for estimating annotator reliability (Dawid and Skene, 1979), uncertainty quantification via MC Dropout and Deep Ensembles (Zhou et al., 2022), and full label distribution modeling (Meissner et al., 2021).

Multitask approaches. Mostafazadeh Davani et al. (2022) proposed a multitask framework with separate classification heads per annotator, enabling individualized predictions while sharing text representations. Jinadu and Ding (2024) extended this with explicit noise tolerance. However, multitask methods suffer from underdetermination for sparse annotators who contribute few labels.

Embedding-based approaches. Recent work represents annotators as learned embeddings integrated early in the network. Deng et al. (2023) jointly learn annotator and annotation embeddings to capture fine-grained subjectivity signals. Mokhberian et al. (2024) introduced AART, which uses contrastive learning (InfoNCE) to align annotator embeddings with labeling patterns, demonstrating more equitable performance for sparse annotators. However, they noted that “differentiating between labeling noise and natural disagreement is a challenge that remains unaddressed,” a gap our work directly targets.

2.3 Learning from Noisy Labels

Noise correction targets errors in labels, such as random flips and annotation mistakes, to improve data quality and robustness (Zhan et al., 2019). Cross-entropy easily fits noisy labels (Zhang et al., 2017), while bootstrapping adds prediction terms to reduce this effect (Reed et al., 2015). Deep models learn clean patterns before memorizing noise (Arazo et al., 2019; Liu et al., 2020; Li et al., 2020; Nishi et al., 2021), a property exploited by many denoising methods. Effective correction must remove misleading labels without discarding true signal (Jinadu et al., 2023; Arazo et al., 2019; Jinadu and Ding, 2024).

3 Preliminary

We first set up the general multi-annotator learning problem. We then discuss two high-level network architecture approaches for modeling multi-annotator datasets. 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 $A = \{a_j\}_{j=1}^J$, and annotations $Y = \{1, \dots, Q\}$. 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 $\{y_{ij}\}_{j=1}^J$.

Multi-task approaches are seen in several previous works (Jinadu et al., 2023; Fornaciari et al., 2021; Mostafazadeh Davani et al., 2022; Jinadu and Ding, 2024), and basically are a generalization of the single-task approach. They train a separate, fully connected layer for each annotator to learn the annotator-specific labeling behavior.

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. 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 embedding of annotator a_j , and $\oplus$ is the fusion operation that we treat as a simple addition in our work. This fused, annotator-aware representation is then fed to a classification model to make a prediction.

A few methods utilize this approach. For example, the approach proposed by Mokhberian et al. (2024) directly incorporates the annotator embeddings into the text representations. We refer to this method as Annotator Aware Representations for Texts (AART) in this paper. Another technique proposed by Deng et al. (2023) additionally incor-

porates annotation embeddings. We refer to this method as **MichEmbed** in this paper.

4 Methodology: Noise-Robust Annotator Embedding (NRA-Embed)

Embedding-based approaches with contrastive loss achieve strong results in the learning from multiple annotator setting (Mokhberian et al., 2024; Deng et al., 2023), as the contrastive objective enables learning from agreement and disagreement patterns across annotators. However, Figure 1 shows that these methods degrade under annotation noise.

RINCE introduces a symmetric adaptation of InfoNCE that downweights corrupted positive and negative pairs, providing robustness to label noise in binary classification (Ghosh et al., 2015; Chuang et al., 2022). In our setting, annotator pairs are treated as positive if they agree on the label of a text instance and negative if they disagree. The RINCE loss is defined as (Chuang et al., 2022):

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

where s^+ is the similarity score (e.g., cosine similarity) for a positive (agreement) pair, s_i^- are scores for negative (disagreement) pairs, and $\lambda > 0$ weights the contribution of negatives. The parameter $q \in (0, 1]$ interpolates between InfoNCE’s expressive power ($q \rightarrow 0$) and RINCE’s noise robustness ($q = 1$); a full theoretical analysis of this trade-off is provided in Appendix E.

To jointly learn task performance and annotator embeddings, we pass combined embeddings $g(x_i, a_j)$ through a classification layer and optimize the following composite objective:

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

The first term is a standard cross-entropy loss predicting annotator a_j ’s label for input x_i . The second is an ℓ_2 regularization term to the annotator embeddings a_j , following Mokhberian et al. (2024); in our experiments we set λ_1 to 0 as described in Section 5.4. The third structures the annotator embedding space via RINCE: label-concordant annotator pairs are encouraged to align while label-discordant pairs are repelled, with RINCE’s symmetric formulation providing robustness when these agreement/disagreement signals are corrupted by noise. Our NRA-Embed framework adopts a similar

embedding-based multi-annotator architecture to AART (Mokhberian et al., 2024), with a critical modification: replacing the conventional InfoNCE objective with RINCE for contrastive learning. This design choice enables a controlled evaluation of robust contrastive learning while preserving architectural consistency and comparability with existing annotator embedding approaches. Importantly, the RINCE loss is architecture-agnostic and can be integrated into any similarity-based annotator embedding framework that relies on pairwise representation comparisons, without requiring changes to the underlying model structure or training pipeline. While our experiments focus on a specific instantiation of this framework, extending RINCE to other annotator-aware architectures and tasks remains an important direction for future work.

5 Experimental Setup

5.1 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).
- **English Perspectivist Irony Corpus (EPIC)** EPIC is a disaggregated dataset for irony detection consisting of 3,000 short social-media conversations represented as *post-reply* pairs collected from Twitter and Reddit, spanning five regional varieties of English (UK, US, Ireland, Australia, India) (Frenda et al., 2023; Multilingual Perspectivist NLU, 2023). We preprocess EPIC into one instance per annotated post-reply pair and the assigned binary label to the reply. We encode each instance’s text item as a two-segment input where the post is segment A, and the reply is segment B using the tokenizer’s native pair encoding (i.e., passing text and text_pair to the tokenizer). Each instance is annotated by 5 annotators in this dataset.
- **HS-Brexit Dataset (HSB)** The HS-Brexit Dataset (Akhtar et al., 2021)

contains Brexit-related tweets annotated for hate speech, aggressiveness, offensiveness, and stereotypes by six annotators from diverse demographic groups. The dataset captures multiple perspectives on abusive language and uses a polarization index to measure annotator disagreement. The HS-Brexit Dataset has 6 crowdworkers per sample, and we use the hate-speech annotations.

5.2 Baseline Models

We compare against the following baseline methods:

- **Multitask:** We follow the approach proposed by (Mostafazadeh Davani et al., 2022) which involves one fully-connected layer for each annotator with a shared RoBERTa model.
- **AART:** We evaluate the approach introduced by (Mokhberian et al., 2024) 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.
- **MichEmbed:** We follow the approach by (Deng et al., 2023), which we refer to as MichEmbed in this work that utilizes annotator embeddings as well as annotation embeddings and a single fully-connected classification layer built off of RoBERTa in our experiments.

5.3 Noise Injection

Although the datasets we are evaluating on, MDA, EPIC, and HSB, are all per-annotator subjective datasets, meaning that they likely have naturally occurring annotation noise, we further inject controlled noise to the annotation labels. We do this to establish proper baselines that isolate how each method degrades as additional noise is introduced. We inject label noise in two separate ways: symmetric label flipping and instance-dependent noise.

5.3.1 Symmetric Noise

Our primary noise injection mechanism is by independently flipping individual annotator labels with a specified probability. Specifically, for each annotator, we assign a noise level (ranging from 0% to 40% in our experiments), and then independently flip each of that annotator’s labels with probability equal to their assigned noise level. For binary

classification tasks, label flips are implemented by inverting the label (i.e., $0 \rightarrow 1$ or $1 \rightarrow 0$). This per-annotator, per-label flipping approach allows us to model annotator-specific noise patterns while maintaining the multi-annotator structure of the data. 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. We also evaluate on the multiclass Sentiment Analysis (SNT) (Díaz et al., 2018) dataset using symmetric noise injection for completeness. Results and discussion are reported in Appendix B. All label noise is injected only in the training set. Test sets remain clean from any injected noise to enable fair evaluation and see how effective the various methods are at filtering out noise while keeping subjective opinions and minority voices.

5.3.2 Instance-Dependent Noise

While symmetric label flipping provides a controlled baseline for measuring robustness, it assumes that every instance is equally susceptible to annotation error. In practice, annotation errors are heterogeneous across instances: semantically ambiguous or confusing inputs are more likely to be mislabeled than the clear-cut ones. To test whether NRA-Embed’s robustness extends beyond uniform noise, we additionally evaluate under *instance-dependent* noise, roughly following the synthetic generation procedure of Xia et al. (2020).

Specifically, we compute a per-instance flip probability d_i by projecting frozen RoBERTa [CLS] embeddings e_i onto a random vector $w \sim \mathcal{N}(0, 0.5)$ with a fixed seed for reproducibility:

$$d_i = \sigma(e_i \cdot w), \quad (3)$$

where σ is the sigmoid function. We then rescale d_i so that the population mean equals the target global noise rate ε :

$$d_i \leftarrow \text{clip} \left(d_i \cdot \frac{\varepsilon}{\bar{d}}, 0, 1 \right), \quad (4)$$

where $\bar{d}$ is the empirical mean of d_i over the training set. For each annotator’s label on each instance, we independently flip the binary label with probability d_i . This procedure yields heterogeneous noise where semantically similar instances have correlated flip probabilities. Unlike Xia et al. (2020), who use a class-conditional projection w_y for multi-class settings, we use a single global w , which is sufficient for the binary subjective tasks considered here. As with symmetric noise, all noise is injected only in the training set.

5.4 Implementation Details

We implemented the classification models using the HuggingFace transformers library (version 5.5.0) (Wolf et al., 2020). Our experimental setup for the annotator embedding approach for subjective classification closely resembles that of Mokherian et al. (2024), where annotator embeddings are initialized from a standard random distribution with the same dimensionality as the model’s hidden layers. For all experiments, we train the models for 10 epochs. 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. The hyperparameters $\lambda_1 = 0$ and $\lambda_2 = 0.1$ were used for all experiments. We set $\lambda_1 = 0$ to isolate the effects of the contrastive loss terms. The q parameter values used for experiments are specified in the results tables; we analyze the effect of q on robustness in Appendix A. All reported results (unless specified otherwise) represent means over 5 independent runs with different random seeds.

5.5 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 based on the **Mean-Annotator F1** defined in Mokherian et al. (2024), which we compute as follows across all J annotators:

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

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

6 Results

We present the main results in Table 1 as mean-annotator F1 scores. This metric weights all annotators equally regardless of contribution volume, ensuring sparse annotator representation. Because Majority Vote produces a single item-level prediction rather than annotator-specific outputs, we

evaluate it using accuracy as well as F1, while all annotator-aware models are evaluated using Mean-Annotator F1. We see that our method performs the best or competitively in many conditions, with the clearest gains at high noise levels.

6.1 Annotator Embedding Approach + Noise Robustness Enhancements

We compare our NRA-Embed approach against several baselines to measure gains in embedding stability under synthetic annotation noise. Table 1 presents mean annotator F1 scores on three benchmark datasets, MDA, EPIC, and HSB, at no noise, 20% noise, 30% noise, and 40% noise. Our Noise-Robust Annotator Embedding Approach consistently improves or performs competitively to the baselines. Cells marked $\ddagger$ in Table 1 denote where *NRA-Embed* is significantly better than every baseline except AART (Welch’s t-test, $N = 5$ seeds, Holm–Bonferroni corrected across 48 comparisons), concentrated in the high-noise (30%, 40%) regime on MDA and EPIC where noise robustness matters most. Differences between *NRA-Embed* and AART fall within seed variance across all 12 cells.

6.2 Robustness to Instance-Dependent Noise

To test whether our robustness findings extend beyond symmetric flips, we evaluate all methods under instance-dependent noise (Section 5.3.2). Table 3 reports mean annotator F1 across MDA, EPIC, and HSB at three global noise rates. The dataset-structure pattern from symmetric noise persists. On MDA and EPIC, contrastive embedding methods substantially outperform multitask and MichEmbed at every noise level, with the gap widening under higher noise: at 40%, NRA-Embed reaches 0.667 on EPIC and 0.741 on MDA, while multitask collapses to 0.582 and 0.444. NRA-Embed and AART remain within seed variance throughout. On HSB, multitask is numerically best at most noise levels, consistent with its advantage on small datasets with dense per-annotator labels. These results indicate that NRA-Embed’s robustness is not specific to the symmetric noise model, and that dataset structure remains a stronger determinant of method choice than noise type.

6.3 Statistical Parity Across Annotator Groups

Mean-annotator F1 weights all annotators equally, but it does not reveal whether a particular anno-

Dataset	Noise Level	Majority Vote	Multitask	MichEmbed	AART	Ours
MDA	No Noise	0.6031 (0.5206)	0.5121 $\pm$ 0.006	0.7867 $\pm$ 0.006	<u>0.7912 $\pm$ 0.005</u>	0.7927 $\pm$ 0.005
	20% Noise	0.6007 (0.5191)	0.4961 $\pm$ 0.018	0.7679 $\pm$ 0.018	<u>0.7902 $\pm$ 0.0033</u>	0.7905 $\pm$ 0.003
	30% Noise	0.6127 (0.5253)	0.4740 $\pm$ 0.013	0.7124 $\pm$ 0.018	<u>0.7814 $\pm$ 0.011</u>	0.7839 $\pm$ 0.01[‡]
	40% Noise	0.5850 (0.5114)	0.4265 $\pm$ 0.035	0.5211 $\pm$ 0.059	<u>0.7183 $\pm$ 0.026</u>	0.7208 $\pm$ 0.03[‡]
EPIC	No Noise	0.6222(0.5306)	0.5131 $\pm$ 0.008	0.7879 $\pm$ 0.007	<u>0.7925 $\pm$ 0.005</u>	0.7935 $\pm$ 0.005
	20% Noise	0.6101(0.5236)	0.4938 $\pm$ 0.023	0.7728 $\pm$ 0.022	<u>0.7881 $\pm$ 0.0015</u>	0.7891 $\pm$ 0.003
	30% Noise	0.5996(0.5177)	0.4741 $\pm$ 0.018	0.7073 $\pm$ 0.017	<u>0.7783 $\pm$ 0.012</u>	0.7792 $\pm$ 0.008[‡]
	40% Noise	0.5767(0.5074)	0.4342 $\pm$ 0.032	0.5116 $\pm$ 0.058	<u>0.7195 $\pm$ 0.027</u>	0.7228 $\pm$ 0.028[‡]
HSB	No Noise	0.8524 (0.7486)	0.9318 $\pm$ 0.004	0.9347 $\pm$ 0.004	0.9304 $\pm$ 0.003	<u>0.9329 $\pm$ 0.004</u>
	20% Noise	0.7776 (0.6509)	0.9134 $\pm$ 0.008	0.9112 $\pm$ 0.011	0.9096 $\pm$ 0.008	<u>0.9112 $\pm$ 0.011</u>
	30% Noise	0.7129 (0.5798)	0.8724 $\pm$ 0.014	<u>0.8783 $\pm$ 0.02</u>	0.8757 $\pm$ 0.03	0.8793 $\pm$ 0.019
	40% Noise	0.6639 (0.5288)	<u>0.7740 $\pm$ 0.03</u>	0.7583 $\pm$ 0.05	0.7505 $\pm$ 0.08	0.7830 $\pm$ 0.082

Table 1: Mean annotator-level F1 across three datasets (MDA, EPIC, and HSB) under varying synthetic label-noise levels (mean $\pm$ std over $N = 5$ seeds). Best is **bold**; second best is underlined. Our method performs best or second best in most conditions, especially under high-noise conditions (The q-value chosen varies depending on the setting that provides the best results, with the majority being $q = 1.0$). In brackets after each F1 in Majority Vote (MV), we report the baseline accuracy computed with the same item-annotator metric. Majority-vote represents the single-task model trained on majority-aggregated labels. For each item x_i , we collapse annotator labels $\{y_{ij}\}_j$ to a single consensus $y_i^{\text{MV}} = \text{mode}(\{y_{ij}\}_j)$ (ties broken deterministically by the training-set class prior), predict y_i^{MV} for all annotators, and score $\frac{|\{(i,j): y_i^{\text{MV}} = y_{ij}\}|}{|\{(i,j)\}|}$. [‡] indicates that *Ours* is significantly better than every baseline except one at the corrected $\alpha = 0.05$ level (Welch’s two-sample t-test from the reported mean $\pm$ std with $N = 5$, Holm-Bonferroni corrected across all 48 pairwise comparisons); the non-significant baseline in every [‡] cell is AART. Differences between *Ours* and AART are within seed variance across all 12 cells.

tator group’s opinions were preserved. Following Mokhberian et al. (2024), we therefore report *statistical parity*: the gap in mean per-annotator F1 between a majority and a minority group of annotators. We define the minority group as the bottom 25th percentile of annotators by similarity to the clean majority-vote label, i.e., the annotators whose judgments most often diverge from consensus. Group membership is computed once from the clean labels and held fixed across noise levels and models, so the groups being compared do not shift as corruption is introduced. We focus on MDA and EPIC, our two large annotator-pool datasets that allow for clear minority groups; a second grouping, ranking annotators by contribution count, is reported in Appendix C. However, it is important to note that a small parity gap is not on its own evidence of fairness, since a method can equalize performance by being uniformly poor across all groups. We therefore report absolute minority-group F1 alongside the gap. Table 2 makes this distinction concrete. Majority Vote attains the smallest gap on both datasets at every noise level (0.031–0.053 on MDA, 0.115–0.154 on

EPIC), but its minority-group F1 trails AART and NRA-Embed by a wide margin: at 40% noise it reaches only 0.516 on MDA and 0.494 on EPIC, against 0.683 and 0.545 for NRA-Embed. Its parity therefore reflects uniformly weak predictions across all annotators. Multitask fails on both axes on EPIC, combining the largest gap of any method (0.271–0.405) with the lowest minority-group F1. AART and NRA-Embed show larger gaps than Majority Vote but the highest minority-group F1 in every cell, meaning they recover minority annotators’ judgments more accurately in absolute terms even where the parity score alone looks less favorable. Both also degrade minimally: NRA-Embed’s minority-group F1 falls from 0.723 to 0.683 on MDA and from 0.633 to 0.545 on EPIC as noise rises from 20% to 40%, whereas MichEmbed drops from 0.692 to 0.472 and from 0.597 to 0.454 over the same range. Most importantly for our central claim, NRA-Embed’s gap and minority-group F1 track AART within seed variance at every noise level on both datasets: the noise robustness of NRA-Embed does not come at the cost of disproportionately worse performance for annotators

Method	Gap ↓ / Minority F1 ↑		
	20%	30%	40%
<i>MDA</i>			
Multitask	0.055/0.472	0.079/0.445	0.066/0.396
MichEmbed	0.097/0.692	0.095/0.643	0.091/0.472
MajVote	0.031 /0.549	0.039 /0.526	0.053 /0.516
AART	0.091/ <u>0.723</u>	0.084/ 0.724	0.082/ <u>0.682</u>
Ours	0.089/ 0.723	0.083/ <u>0.721</u>	0.078/ 0.683
<i>EPIC</i>			
Multitask	0.405/0.403	0.383/0.401	0.271/0.380
MichEmbed	0.178/0.597	0.151/0.548	0.149/0.454
MajVote	0.154 /0.536	0.142 /0.504	0.115 /0.494
AART	0.203/0.630	0.192/0.611	0.159/ 0.547
Ours	0.201/ 0.633	0.193/ 0.611	0.165/ <u>0.545</u>

Table 2: Statistical parity under instance-dependent label noise on MDA and EPIC (mean over $N = 5$ seeds). Each cell reports *parity gap* / *minority-group F1*. Following Mokherian et al. (2024), the minority group is the bottom 25th percentile of annotators by similarity to the clean majority-vote label. The gap is $|\text{majority-group F1} - \text{minority-group F1}|$ (lower is better); minority F1 is absolute (higher is better). Best per column is **bold**, second best underlined, scored separately for each of the two quantities. Standard deviations and a second annotator grouping are in Appendix C.

who disagree with the majority.

6.4 Impact of Renegade Annotators on Model Performance

We also 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. Results demonstrate that no method consistently outperforms the others under conditions of renegade annotation. This suggests that robustness to systematic adversarial annotators may require explicit detection or weighting mechanisms beyond contrastive robustness. Detailed performance metrics are provided in Table 4.

7 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.

Contrastive embedding methods are the stronger choice for sparse-annotator settings. Across both noise models, AART and NRA-Embed consistently outperform multitask and MichEmbed

on the large sparse-annotator datasets (MDA, EPIC). Within this family, RINCE serves as a low-cost refinement over InfoNCE: it matches AART in clean settings while providing increasing numerical advantages as noise grows, positioning NRA-Embed as a drop-in robustness enhancement for embedding-based annotator models.

Annotator embedding models outperform multitask learning. Our results show that models such as MichEmbed, AART, and our approach NRA-Embed, which learn annotator embeddings, generally outperform multitask approaches, especially at high noise. 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. We further analyze per-annotator performance across methods on HSB in Appendix D.

Multitask models degrade in performance with many, sparse annotators. The multitask model performed the worst of the annotator-aware models for the SNT dataset (Appendix B). 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 5). This aligns with Mokherian et al. (2024), who showed through statistical parity analysis that multitask models exhibit the highest performance disparity for sparse annotators, whereas embedding-based approaches give more balanced and fair results. On the other hand, the multitask model performs well on HSB which has a smaller amount of annotators, who each labeled many samples.

Binary classification aligns with contrastive objectives. Our main experiments focus on binary tasks where contrastive agree/disagree signals match the label structure. Supplementary experiments on the multiclass Likert-scale SNT dataset (Appendix B) show that contrastive methods underperform, suggesting ordinal settings require class-sensitive modifications.

8 Conclusion

We explored the distinction between 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 ac-

Dataset	Noise Level	Multitask	MichEmbed	AART	Ours
MDA	20%	0.5085 $\pm$ 0.022	0.7634 $\pm$ 0.015	0.7895 $\pm$ 0.011	0.7874 $\pm$ 0.011
	30%	0.5012 $\pm$ 0.022	0.7129 $\pm$ 0.035	0.7851 $\pm$ 0.011	0.7821 $\pm$ 0.011
	40%	0.4443 $\pm$ 0.019	0.5387 $\pm$ 0.082	0.7417 $\pm$ 0.013	0.7406 $\pm$ 0.005
EPIC	20%	0.7039 $\pm$ 0.010	0.7293 $\pm$ 0.029	0.7814 $\pm$ 0.005	0.7824 $\pm$ 0.004
	30%	0.6851 $\pm$ 0.028	0.6600 $\pm$ 0.052	<u>0.7529 $\pm$ 0.015</u>	0.7542 $\pm$ 0.015
	40%	0.5815 $\pm$ 0.038	0.5651 $\pm$ 0.050	<u>0.6653 $\pm$ 0.042</u>	0.6671 $\pm$ 0.045
HSB	20%	0.9110 $\pm$ 0.011	0.9011 $\pm$ 0.015	0.9034 $\pm$ 0.005	0.9002 $\pm$ 0.011
	30%	0.8499 $\pm$ 0.027	0.8551 $\pm$ 0.019	0.8465 $\pm$ 0.034	0.8441 $\pm$ 0.038
	40%	0.7777 $\pm$ 0.049	<u>0.7423 $\pm$ 0.067</u>	0.7097 $\pm$ 0.089	0.7159 $\pm$ 0.081

Table 3: Mean annotator-level F1 under **instance-dependent** label noise across three datasets (MDA, EPIC, HSB), at varying global noise rates ϵ (mean $\pm$ std over $N = 5$ seeds). Following (Xia et al., 2020), per-instance flip probabilities are derived from a random projection of RoBERTa embeddings and rescaled to match the target global noise rate. Best is **bold**; second best is underlined. The 0% condition is identical to Table 1 and omitted for brevity. Welch’s t-tests with Holm–Bonferroni correction (27 comparisons) show no significant difference between *Ours* and AART in any cell; *Ours* significantly outperforms Multitask on MDA at every noise level and on EPIC at 20%, and no method is significantly better than any other on HSB.

Model	MDA	EPIC	HSB
Multitask	0.6111	0.7179	0.8002
MichEmbed	0.7779	0.7740	0.8873
AART	<u>0.7755</u>	<u>0.7765</u>	<u>0.8868</u>
NRA-Embed	0.7728	0.7774	0.8695

Table 4: Mean Annotator F1 Scores comparing model robustness to renegade annotators: 10% of annotators have 70% of their annotations corrupted ($N = 5$ seeds). Best is **bold**; second best is underlined. No method consistently dominates across datasets, with most pairwise differences within seed variance.

DS	#A	#E/#A	#CW	#S	#L
MDA	819	60	5	44k	2
EPIC	74	192	5	14.7k	2
HSB	6	952	6	5.7k	2
SNT	1481	41	4–12	60.4k	5

Table 5: Dataset statistics. #A is the number of annotators, #E/#A is the average number of examples per annotator, #CW is the number of annotators per example, #S is the number of sample annotations, and #L is the number of labels.

tivities 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 suggest that embedding-based approaches are the more reliable methodology for training in multi-annotator cases with many, sparse annotators. Furthermore, we recommend that dataset raw labels should be released,

however noisy, so that issues with label noise can be directly addressed by the model.

9 Limitations

Our evaluation relies on two forms of synthetic noise injection (symmetric and instance dependent), which may not fully capture the complexity of real-world annotation errors. Future work should explore noise models that better reflect actual annotator behavior, such as systematic biases, attention lapses, or task-specific confusion patterns.

Per-annotator modeling becomes computationally expensive in datasets with many annotators. For large-scale applications, future work should investigate annotator grouping strategies or approximation methods to reduce computational overhead while preserving perspective diversity.

Our approach depends on existing datasets with multiple annotations per sample. Like any crowd-sourced dataset, these have inherent limitations in representing population diversity due to cost constraints (typically 3-5 annotators per sample), annotator recruitment biases, and geographic skew. While our method robustly models available perspectives, it cannot capture viewpoints absent from the annotation pool. Future work should explore methods to incorporate high-quality minority perspectives, such as targeted recruitment or active learning strategies that identify underrepresented groups. In addition, our evaluation focuses on mean annotator F1, which captures per-annotator performance averaged across a test set. We do not directly evaluate disagreement at the instance level, which would help further support the claim of minority voice preservation. However, we do

evaluate statistical parity across minority groups. Applying metrics for item-level disagreement such as those proposed in (Mokhberian et al., 2024), is non-trivial as item level disagreement would be further affected by the injected corruption. Future work should explore metrics that account for this noise in the instance disagreement setting.

10 Ethical Considerations

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.

Some content in the HSB and MDA datasets may be considered offensive, so examples of these instances were not included in this manuscript. In addition, we utilized the assistance of AI (LLMs) in some research, coding, and writing.

11 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.

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.
- 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.
- 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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.
- 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. Association for Computing Machinery.
- 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.
- Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. [EPIC: Multi-perspective annotation of a corpus of irony](#). In *Proceedings*

- of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.
- Aritra Ghosh, Naresh Manwani, and PS Sastry. 2015. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pages 8535–8545.
- 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.
- Uthman Jinadu, Jesse Annan, Shanshan Wen, and Yi Ding. 2023. Loss modeling for multi-annotator datasets. In *NeurIPS 2023 Workshop on Algorithmic Fairness through the Lens of Time*. Spotlight paper.
- 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.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*.
- 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. DeSMOG: Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.
- Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.
- Negar Mokherian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.
- 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.
- Multilingual Perspectivist NLU. 2023. [Multilingual-Perspectivist-NLU/EPIC \(EPICorpus\) dataset card](#). Hugging Face Datasets. Accessed 2026-01-04.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9131–9143, Online. Association for Computational Linguistics.
- 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 Salieb-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](#). In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pages 2142–2151.
- 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, Berlin, Germany. Association for Computational Linguistics.
- 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. [Unifying data perspectivism and personalization: An application to social norms](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations (Workshop Track)*.
- 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. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- 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.
- Zeera 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.
- Zeera 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.
- Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. 2020. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, pages 7597–7610.
- 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. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.
- Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

A Analysis on q Parameter

A.1 Impact of Parameter q on Noise-Robustness in Annotator Embedding

Table 6 examines the effect of the robustness parameter q across noise levels. The parameter q controls the trade-off between learning from hard samples (low q , closer to standard InfoNCE) and robustness to noisy labels (high q). We observe that $q=1.0$ consistently achieves the highest or near-highest numerical performance at 40% noise across all datasets. This aligns with our intuition that stronger robustness helps under severe label corruption. At moderate noise levels (20%), the pattern is less consistent, with optimal q varying by dataset. However, pairwise comparisons across q values do not reach statistical significance (Welch’s t-test, $p > 0.05$ for all comparisons, $n=5$ runs per configuration). This suggests that while $q=1.0$ offers a slight numerical advantage at high noise, performance is generally stable across q values. In practice, this is useful, as practitioners can default to $q=1.0$ for annotation settings they expect to be noisy without requiring extensive hyperparameter tuning.

B Supplementary Results on Multiclass Data

To assess generalization beyond binary classification, we evaluated on the Sentiment Analysis Dataset (SNT), introduced by Díaz et al. (2018). SNT 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.” The dataset uses a 5-class Likert scale and contains 1,481 annotators averaging 41 examples each, with 4–12 annotators per sample (see Table 5).

B.1 Noise Injection for Multiclass Labels

For SNT, we applied symmetric noise injection following a similar procedure to the binary datasets. Each annotator’s labels were flipped independently according to the noise percentage, to one of the four other classes. As with our binary experiments, all noise was injected only in the training set; test labels remained clean.

B.2 Discussion: Limitations of Contrastive Loss for Likert-Scale Data

Our main experiments focus on binary classification tasks, where the contrastive loss’s

agree/disagree signal naturally aligns with the label structure. The SNT results reveal important limitations when extending contrastive approaches to multiclass ordinal data.

Likert scales degrade contrastive loss performance. On SNT, which uses Likert scale classification, we find that previously strong contrastive approaches like AART and our NRA-Embed method degrade in performance relative to MichEmbed. We hypothesize this is due to the fundamental mismatch between contrastive objectives and ordinal label structure. A contrastive loss treats labels “Strongly Agree” and “Moderately Agree” as a negative pair in exactly the same way it treats “Strongly Agree” and “Strongly Disagree”—both are simply instances of disagreement. This binary treatment discards the semantic similarity between adjacent ordinal categories, leading to suboptimal representation learning.

In contrast, MichEmbed, which relies on a combination of Annotator and Annotation Embeddings rather than contrastive loss, performs strongly on SNT. The annotation embeddings can implicitly capture relationships between label values, making this approach better suited for ordinal classification.

Multitask models struggle with sparse annotators. The multitask model performed particularly poorly on SNT (Table 7). This is likely due to the large number of annotators (1,481) relative to their average annotation count (41 samples each), creating extremely sparse per-annotator training data (see Table 5). In contrast, multitask performed well on HSB, which has only 6 annotators who each labeled approximately 952 samples. This pattern aligns with findings from Mokherberian et al. (2024), who demonstrated that embedding-based approaches provide more equitable performance for sparse annotators.

Implications for future work. These findings suggest that extending noise-robust contrastive learning to multiclass ordinal settings requires class-sensitive modifications. Future work should explore ordinal-aware contrastive objectives that weight negative pairs by label distance, for instance, penalizing “Strongly Agree” vs. “Strongly Disagree” more heavily than “Strongly Agree” vs. “Moderately Agree.” Formally, the objective in Eq. 2 partitions annotator pairs by the indicator $1[y_{ij} = y_{ij'}]$, which discards the distance structure

	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	0.7915 ± 0.005	0.7906 ± 0.005	0.7927 ± 0.004	0.7891 ± 0.003	0.7905 ± 0.003	0.7901 ± 0.004	0.7118 ± 0.043	0.7099 ± 0.037	0.7174 ± 0.027
HSB	0.9315 ± 0.003	0.9302 ± 0.005	0.9321 ± 0.004	0.9112 ± 0.011	0.9070 ± 0.009	0.9028 ± 0.016	0.7635 ± 0.096	0.7678 ± 0.100	0.7782 ± 0.110
EPIC	0.7926 ± 0.006	0.7913 ± 0.006	0.7935 ± 0.005	0.7869 ± 0.0022	0.7891 ± 0.003	0.7883 ± 0.0028	0.7207 ± 0.04	0.7167 ± 0.027	0.7228 ± 0.028

Table 6: Mean Annotator F1 Scores (top row) with standard deviations (bottom row) showing the effect of q on robustness across noise levels (Seeds are taken from earlier runs than Table 1 resulting in slightly different scores).

Dataset	Noise	MV	Multitask	MichEmbed	AART	Ours
SNT	0%	0.245	0.267	<u>0.517</u>	0.521	0.494
	10%	0.242	0.268	<u>0.492</u>	0.500	0.489
	20%	0.251	0.259	0.483	<u>0.475</u>	0.471
	30%	0.235	0.248	0.446	<u>0.439</u>	0.441
	40%	0.236	0.225	0.418	<u>0.403</u>	0.406

Table 7: Mean annotator-level F1 on the multiclass SNT dataset under varying noise levels (symmetric). Best is **bold**; second best is underlined. Unlike binary datasets, contrastive methods (AART, NRA-Embed) do not consistently outperform MichEmbed. MV = Majority Vote baseline.

	No Noise			20% Noise			40% Noise		
q	0.5	0.75	1.0	0.5	0.75	1.0	0.5	0.75	1.0
SNT	0.494 ± 0.003	0.495 ± 0.005	0.502 ± 0.005	0.471 ± 0.013	0.463 ± 0.020	0.464 ± 0.011	0.400 ± 0.012	0.396 ± 0.019	0.400 ± 0.017

Table 8: Effect of robustness parameter q on SNT. Unlike binary datasets, the pattern is inconsistent: $q = 0.5$ performs best at 20% noise, while $q = 1.0$ is marginally better at both 0% and 40% noise. This analysis’s runs differ from Table 7’s seeds resulting in slightly different scores.

of an ordinal scale: annotators one Likert point apart are repelled as strongly as those at opposite ends. A natural fix is to weight pairs by normalized ordinal distance $\delta_{jj'} = |y_{ij} - y_{ij'}|/(Q - 1)$, treating them as soft positives with weight $(1 - \delta_{jj'})$ and soft negatives with weight $\delta_{jj'}$. Under RINCE this scales each negative’s contribution to the partition function by $\delta_{jj'}$, preserving the gradient-suppression mechanism of Appendix E while respecting the label metric.

C Statistical Parity: Full Results

Table 9 reports the standard deviations omitted from Table 2, and Table 10 reports a second an-

notator grouping in which the minority group is the bottom 25th percentile by contribution count (number of items labeled), following the sparse-annotator analysis of Mokherian et al. (2024). On EPIC this grouping behaves as expected: low-count annotators are predicted less accurately than high-count annotators for all of the methods. On MDA the direction is the opposite, with the low-count group scoring *higher* than the majority group for every method at every noise level. Because statistical parity is an absolute difference, the MDA gaps in Table 10 should not be read as evidence of disadvantage for sparse annotators on that dataset. Under both groupings, NRA-Embed and AART remain within seed variance of each other at every noise level.

D Per-Annotator Analysis (HSB)

To examine how different models capture individual annotator perspectives, we conduct a detailed analysis on the HSB dataset, which contains annotations from six crowdworkers (A1–A6) with diverse labeling patterns. Table 11 presents F1 scores for each annotator across all methods and noise conditions.

D.1 Annotator Heterogeneity

Even under clean conditions, substantial performance variation exists across annotators. We observe two distinct groups: *high-agreement annotators* (A1, A2, A3) who achieve F1 scores of 0.95 or above with embedding-based methods, and *low-agreement annotators* (A4, A5, A6) who score at or below 0.93. Notably, A5 consistently exhibits the lowest F1 across all methods (0.88–0.89 at 0% noise), suggesting this annotator’s labeling patterns diverge most from learnable representations. This

Method	20% Noise		30% Noise		40% Noise	
	Parity Gap ↓	Minority F1 ↑	Parity Gap ↓	Minority F1 ↑	Parity Gap ↓	Minority F1 ↑
<i>MDA</i>						
Multitask	0.055 ±0.039	0.472 ±0.021	0.079 ±0.061	0.445 ±0.029	0.066 ±0.053	0.396 ±0.033
MichEmbed	0.097 ±0.023	0.692 ±0.022	0.095 ±0.020	0.643 ±0.036	0.091 ±0.049	0.472 ±0.089
Majority Vote	0.031 ±0.009	0.549 ±0.047	0.039 ±0.017	0.526 ±0.044	0.053 ±0.024	0.516 ±0.038
AART	0.091 ±0.010	0.723 ±0.008	0.084 ±0.011	0.724 ±0.008	0.082 ±0.029	0.682 ±0.017
NRA-Embed (Ours)	0.089 ±0.008	0.723 ±0.006	0.083 ±0.015	0.721 ±0.009	0.078 ±0.027	0.683 ±0.018
<i>EPIC</i>						
Multitask	0.405 ±0.027	0.403 ±0.026	0.383 ±0.055	0.401 ±0.055	0.271 ±0.070	0.380 ±0.050
MichEmbed	0.178 ±0.028	0.597 ±0.033	0.151 ±0.033	0.548 ±0.060	0.149 ±0.052	0.454 ±0.060
Majority Vote	0.154 ±0.020	0.536 ±0.018	0.142 ±0.027	0.504 ±0.028	0.115 ±0.024	0.494 ±0.028
AART	0.203 ±0.011	0.630 ±0.006	0.192 ±0.027	0.611 ±0.020	0.159 ±0.050	0.547 ±0.036
NRA-Embed (Ours)	0.201 ±0.010	0.633 ±0.005	0.193 ±0.028	0.611 ±0.021	0.165 ±0.042	0.545 ±0.039

Table 9: Table 2 with standard deviations over $N = 5$ seeds. Minority group defined by similarity to the clean majority-vote label (bottom 25th percentile).

Method	20% Noise		30% Noise		40% Noise	
	Parity Gap ↓	Minority F1 ↑	Parity Gap ↓	Minority F1 ↑	Parity Gap ↓	Minority F1 ↑
<i>MDA</i>						
Multitask	0.087 ±0.032	0.571 ±0.036	0.118 ±0.020	0.586 ±0.025	0.054 ±0.053	0.483 ±0.051
MichEmbed	0.123 ±0.026	0.852 ±0.012	0.134 ±0.056	0.809 ±0.042	0.099 ±0.049	0.606 ±0.073
Majority Vote	0.081 ±0.019	0.630 ±0.059	0.076 ±0.040	0.599 ±0.083	0.070 ±0.029	0.589 ±0.076
AART	0.117 ±0.009	0.874 ±0.008	0.126 ±0.021	0.876 ±0.019	0.102 ±0.010	0.816 ±0.014
NRA-Embed (Ours)	0.116 ±0.008	0.871 ±0.009	0.122 ±0.017	0.870 ±0.015	0.087 ±0.016	0.803 ±0.015
<i>EPIC</i>						
Multitask	0.191 ±0.047	0.562 ±0.040	0.182 ±0.063	0.550 ±0.067	0.177 ±0.094	0.450 ±0.081
MichEmbed	0.086 ±0.026	0.665 ±0.036	0.075 ±0.040	0.604 ±0.067	0.083 ±0.077	0.504 ±0.089
Majority Vote	0.054 ±0.026	0.611 ±0.020	0.048 ±0.024	0.581 ±0.037	0.039 ±0.032	0.555 ±0.034
AART	0.079 ±0.014	0.723 ±0.012	0.056 ±0.015	0.711 ±0.013	0.023 ±0.015	0.648 ±0.044
NRA-Embed (Ours)	0.080 ±0.008	0.723 ±0.009	0.060 ±0.011	0.710 ±0.016	0.025 ±0.018	0.648 ±0.049

Table 10: Statistical parity with the minority group defined as the bottom 25th percentile of annotators by contribution count (number of items labeled), mean $\pm$ std over $N = 5$ seeds. On MDA the low-count group scores *higher* than the majority group for every method and noise level, so gaps on that dataset do not indicate disadvantage; on EPIC the low-count group scores lower, as expected.

heterogeneity reflects genuine differences in annotator behavior rather than random variation, as the ranking remains stable across methods.

D.2 Cross-Annotator Equity

The Gap metric (Max – Min F1) measures how equitably a model performs across annotators. At 0% noise, NRA-Embed achieves the smallest Gap (0.07), followed by AART (0.08), indicating that contrastive embedding approaches produce more balanced representations than Multitask (0.09) or MichEmbed (0.10). This advantage persists at 20% noise, where NRA-Embed maintains a Gap of 0.08 while Multitask’s Gap increases sharply to 0.15. The widening disparity for Multitask suggests that per-annotator classification heads are more susceptible to noise-induced overfitting for low-agreement annotators.

D.3 Performance Under Noise

At 40% noise, the pattern shifts. While NRA-Embed achieves the second-highest Mean F1 (0.78) among all methods, outperforming AART (0.76) and MichEmbed (0.75), its Gap increases to 0.16. Examining individual annotators reveals why: A5, the hardest annotator, degrades from 0.89 to 0.69 for NRA-Embed ($\Delta = -0.20$), whereas Multitask shows smaller degradation for A5 ($\Delta = -0.15$). However, NRA-Embed shows the smallest degradation for A3 ($\Delta = -0.11$), the annotator where it achieves the best 40%-noise performance (0.85). This suggests that robust contrastive learning helps preserve representations for annotators whose patterns align well with the learned embedding space, but provides less protection for outlier annotators under extreme noise.

Table 11: Per-annotator F1 scores on the HSB dataset (3 runs). **Bold** indicates best performance per annotator at each noise level. Mean averages across annotators; Gap = Max – Min F1 (lower indicates more equitable performance).

	0% Noise				20% Noise				40% Noise			
	Multi	MichE	AART	NRA	Multi	MichE	AART	NRA	Multi	MichE	AART	NRA
A1	.97	.98	.96	.96	.97	.95	.94	.93	.84	.81	.79	.80
A2	.97	.96	.95	.95	.96	.95	.94	.93	.82	.78	.77	.80
A3	.97	.97	.96	.96	.97	.96	.94	.94	.82	.81	.82	.85
A4	.89	.90	.89	.89	.86	.86	.87	.86	.76	.69	.75	.77
A5	.88	.88	.89	.89	.82	.85	.85	.85	.73	.67	.66	.69
A6	.92	.93	.92	.93	.89	.91	.90	.90	.76	.70	.75	.77
Mean	.93	.94	.93	.93	.91	.91	.91	.90	.79	.75	.76	.78
Gap	.09	.10	.08	.07	.15	.11	.10	.08	.11	.14	.16	.16

D.4 Summary

This analysis reveals that: (1) annotator difficulty is consistent across methods, with A5 being hardest and A1–A3 being easiest; (2) NRA-Embed provides the most equitable cross-annotator performance at low-to-moderate noise levels; and (3) at extreme noise, a trade-off emerges between overall performance and cross-annotator equity. The performance of multitask at 40% noise may be attributed to the lack of sparse annotators that is unique to HSB (Akhtar et al., 2021).

E Theoretical Analysis of Robustness in RINCE

In this appendix, we provide the theoretical justification for utilizing the Robust InfoNCE (RINCE) loss (Chuang et al., 2022) in the context of subjective NLP tasks. We analyze and contrast the gradient dynamics of RINCE against the standard InfoNCE loss to demonstrate how the hyperparameter q controls the trade-off between learning from hard samples (exploration) and robustness to label noise (exploitation), and discuss why it’s a perfect fit for our problem requirements. For more details and the theoretical proofs, we refer to RINCE’s original paper (Chuang et al., 2022).

E.1 Formulation

Let s^+ denote the similarity score of a positive pair and $\{s_i^-\}_{i=1}^K$ denote the set of K negative similarity scores. We define the partition function $Z = e^{s^+} + \sum_{i=1}^K e^{s_i^-}$.

The standard **InfoNCE** loss (Oord et al., 2018) is defined as:

$$\mathcal{L}_{\text{InfoNCE}}(s) = -\log\left(\frac{e^{s^+}}{Z}\right) = -s^+ + \log Z \quad (5)$$

The **RINCE** loss utilizes a generalized Box-Cox transformation. For parameters $q \in (0, 1]$ and $\lambda > 0$, the loss is defined as:

$$\mathcal{L}_{\text{RINCE}}^{\lambda, q}(s) = -\frac{e^{qs^+}}{q} + \frac{(\lambda Z)^q}{q} \quad (6)$$

Note that as $q \rightarrow 0$, Eq. (6) asymptotically recovers the behavior of the InfoNCE loss (up to an additive constant).

E.2 Gradient Analysis and Noise Robustness

To understand the robustness mechanism, we analyze the gradient of the loss with respect to the positive score s^+ . This gradient dictates how strongly the model pushes positive pairs together.

InfoNCE Dynamics ($q \rightarrow 0$). The gradient of the standard InfoNCE loss with respect to s^+ is:

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial s^+} = -(1 - p), \quad \text{where } p = \frac{e^{s^+}}{Z} \quad (7)$$

Here, the gradient magnitude is $|1 - p|$. This creates a “hard-positive mining” effect:

- When the model is confident ($p \approx 1$), the gradient is near 0.
- When the model is incorrect ($p \approx 0$), the gradient magnitude is maximized (≈ 1).

While beneficial for clean data, this dynamic is detrimental for noisy or subjective NLP tasks. If a sample is mislabeled or highly subjective (a “noisy positive”), p will be small. InfoNCE will generate a large gradient, forcing the model to overfit this noise.

RINCE Dynamics ($q > 0$). Differentiating Eq. (6) with respect to s^+ yields:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{RINCE}}}{\partial s^+} &= -e^{qs^+} + \lambda^q Z^{q-1} e^{s^+} \\ &= -e^{qs^+} \left(1 - \lambda^q \left(\frac{e^{s^+}}{Z} \right)^{1-q} \right) \quad (8) \\ &= - \underbrace{e^{qs^+}}_{\text{Weighting Term}} \cdot \underbrace{\left(1 - \lambda^q p^{1-q} \right)}_{\text{Error Term}}\end{aligned}$$

The critical innovation in RINCE is the **weighting term** e^{qs^+} . Unlike InfoNCE, the gradient magnitude in RINCE is dependent on the absolute value of the similarity score s^+ :

- If s^+ is large (high confidence), e^{qs^+} is large, maintaining the learning signal.
- If s^+ is small (low confidence/noisy sample), e^{qs^+} shrinks.

Consequently, RINCE applies a “soft clipping” to the gradients. If a sample is too difficult (implying potential noise or extreme subjectivity), the e^{qs^+} term suppresses the gradient, effectively ignoring the outlier rather than corrupting the representation space.

E.3 The Exploration-Exploitation Trade-off

The parameter q governs a fundamental trade-off between exploration and exploitation, which is particularly relevant for subjective NLP:

1. **Exploration** ($q \rightarrow 0$, **InfoNCE**): The loss focuses on “hard” samples (tail of the distribution). This pushes the model to explore boundaries and learn fine-grained distinctions. In subjective tasks, hard samples often arise from annotator disagreement or label noise. Excessive exploration here can lead to overfitting to the noise.
2. **Exploitation** ($q \rightarrow 1$, **Robustness**): As q increases, the weighting term e^{qs^+} dominates. The model focuses on “easy” samples where it is already confident (head of the distribution). This represents *exploitation* of clear, agreed-upon signals while discarding ambiguous data points.

E.4 Suitability for Noisy, Subjective NLP

Our problem setting involves inherently noisy, subjective NLP labels (e.g., sentiment, emotion,

stance), where annotators frequently disagree and supervision is inconsistent. Contrastive pairs derived from such data are plagued by three distinct types of noise: (1) **False Positives** (semantically distinct examples labeled as similar), (2) **False Negatives** (semantically identical examples labeled as different), and (3) **Subjective Variance**. RINCE addresses these specific challenges through the following mechanisms:

- **Robustness to Noisy Positives via Exploitation.**

Subjective datasets naturally produce weak or contradictory positives. As shown in the gradient analysis, the term e^{qs^+} acts as an automated confidence gate. When the model encounters a “noisy positive” (a pair labeled similar but with low semantic alignment s^+), the gradient is down-weighted. This implicit *easy-positive mining* prevents the model from overfitting to specific annotator errors.

- **Bounded Penalty for False Negatives.**

In standard InfoNCE, a False Negative (a positive pair mislabeled as negative) is treated as a “hard negative,” generating a gradient that grows without bound as the model tries to correct the “error” by pushing the pair apart. This can destroy the semantic structure. RINCE, via the power function transformation (x^q), bounds the penalty for hard negatives. This ensures that while the model separates distinct classes, it remains stable in the presence of contradictory negative labels.

- **Geometric Robustness via Wasserstein Distance.**

Theoretically, RINCE optimizes a contrastive lower bound on mutual information derived from the Wasserstein distance measure. This formulation ensures that the learning objective is focused on the global distributional alignment between representations and labels, rather than sample-wise exact matching. In subjective tasks, where individual labels are noisy samples from a latent sentiment distribution, this geometric robustness is crucial for capturing the underlying data manifold.

- **Implicit soft reweighting vs. explicit noise estimation.**

Unlike methods that require estimating a noise transition matrix or setting a hard “forget

rate” based on a known noise ratio (e.g., Co-teaching; [Han et al., 2018](#)), RINCE handles noise implicitly via gradient suppression. While the hyperparameter q controls the degree of robustness, it modulates the curvature of the loss rather than setting a hard threshold, allowing the model to dynamically down-weight samples based on their individual learning difficulty (s^+, s^-) rather than a global noise statistic.