# 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 multi-annotator 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. Through extensive experiments, we demonstrate that NRA-Embed effectively captures subjective disagreement while maintaining robustness to noise, achieving performance that matches or exceeds state-of-the-art methods.

## 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 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.

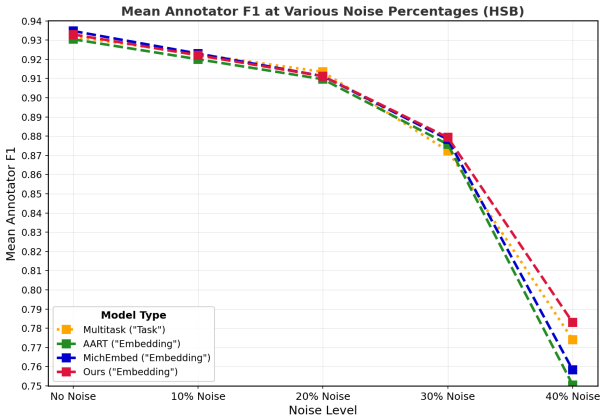


Figure 1: Comparison of Mean Annotator F1 scores for different multi-annotator modeling approaches on the HSB (Akhtar et al., 2021) Dataset under increasing noise levels. We highlight that our embedding-based method demonstrates greater robustness, maintaining higher performance compared to other approaches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

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 (Reidsma and Carletta, 2008), large-scale disagreement learning (Uma et al., 2022), and distributional supervision (Uma et al., 2021) provide principled alternatives to label aggregation.

Building on these foundations, recent methods have moved beyond label aggregation to explicitly model annotator perspectives. These include multi-task architectures with per-annotator classification heads (Davani et al., 2022a; Jinadu and Ding, 2024) and embedding-based approaches that represent annotators as learned vectors (Mokhberian et al., 2023; 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., 2023).

However, contrastive objectives like InfoNCE fundamentally rely on correct positive/negative pair construction, 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 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 embedding framework, NRA-Embed preserves the benefits of contrastive annotator modeling while substantially 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, achieving several significant performance improvements such as a 3.25% increase over AART (Mokhberian et al., 2023) on the HSB (Akhtar et al., 2021) dataset.
- We benchmark across three binary and one multiclass datasets and two noise types (symmetric and renegade), 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 (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., 2021; 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., 2021), and full label distribution modeling (Meissner et al., 2021).

**Multitask approaches.** Davani et al. (2022a) 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 under determination 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. (2023) 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., 2016), while bootstrapping adds prediction terms to reduce this effect (Reed et al., 2014). 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 (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 $A = \{a_j\}_{j=1}^{m}$, 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 $\{y_{ij}\}_{j=1}^{M}$.
***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.

### 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 it 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., 2023) 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 is (Deng et al., 2023), which additionally incorporates annotation embeddings. We refer to this method as **MichEmbed** in this paper.

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

Recent SOTA studies (Mokhberian et al., 2023), Deng et al. (2023) demonstrated that embedding-

based approaches have outstanding potential for multi-annotator classification. Additionally, Mokhberian et al. (2023), showed that contrastive loss enables learning from similarities and dissimilarities among annotators' label choices, a capability absent in multitask approaches that optimize separate heads independently. However, our experiments (depicted in Figure 1) demonstrate that the performance of these methods is degraded in the presence of noisy annotations. The underlying reason is that conventional contrastive losses, such as InfoNCE (Oord et al., 2018; Chen et al., 2020), often fail to learn accurate embeddings that represent annotators' opinions in noisy environments, since 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 that is robust to annotation noise and tunable to emphasize confident and informative annotation signals while down-weighting uncertain or potentially noisy ones. Motivated by this, we propose utilizing Robust InfoNCE (RINCE) (Chuang et al., 2022) and adapting it to our problem setting.

RINCE builds on the insight that contrastive learning with noisy representations can be interpreted as a binary classification with noisy labels over pairwise views (in image context), 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. This positive/negative dichotomy justifies our focus on binary classification datasets, enabling clear alignment between label agreement, disagreement, and contrastive objectives.Ghosh et al. (2015) demonstrates that *symmetric* loss functions offer robustness to label noise in binary classification tasks. RINCE introduces a *symmetric adaptation of InfoNCE* 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 (similarity measure like cosine similarity) 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 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 composite objective function:

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

The first term, $\mathcal{L}_{\text{CE}}$, is a standard cross-entropy loss used to predict annotator $a_j$'s label assignment for input $x_i$ from the joint representation $g(x_i, a_j)$. The second term applies an $\ell_2$ regularization penalty to the annotator embeddings $f(a_j)$ to mitigate overfitting and promote generalization. The third term incorporates the RINCE contrastive loss to structure the annotator embedding space: label-concordant annotator pairs $(a_j, aj')$ are treated as positives and encouraged to align, while label-discordant pairs are repelled. This contrastive objective enforces consistency among reliable annotators while remaining robust to annotation noise.

Our NRA-Embed framework adopts a similar

embedding-based multi-annotator architecture to AART (Mokhberian et al., 2023), 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. We provide a theoretical analysis of RINCE's robustness properties—including its gradient dynamics and the $q$-parameterized exploration–exploitation trade-off—in Appendix C.

# 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; epi). 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.

## 5.2 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.

- **MichEmbed:** We follow the approach by (Deng et al., 2023) which 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

For MDA, EPIC, and HSB, we inject label noise 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 evaluated 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 to enable fair evaluation and see how effective the various methods are at filtering out noise while keeping subjective opinions and minority voices.

### 5.4 Implementation Details

We implemented the classification models using the HuggingFace transformers library (version 4.39) (Wolf et al., 2020). Our experimental setup for the annotator embedding approach for subjective classification closely resembled that of Mokhberian et al. (2023) where Annotator embeddings are initialized from a standard random distribution with the same dimensionality as the model's hidden layers. For all dataset experiments, we trained the models for 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. The hyperparameters $\lambda_1$ = 0.1 and $\lambda_2$ = 0.1 were used for all experiments. The $q$ parameter values used for each experiment are specified in the results tables. All reported results (unless specified otherwise) represent means over 5 independent runs with different random seeds, with standard deviations reported in tables.

### 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 via the **Mean-Annotator F1** (Mokhberian et al. (2023)), 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 $x_i$.

## 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-Aware 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.

### 6.2 Impact of Parameter $q$ on Noise-Robustness in Annotator Embedding

Table 2 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.

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

Table 1: Mean annotator-level F1 across three datasets (MDA, HSB, and EPIC) under varying synthetic label–noise levels. 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 what provides the best results). In brackets after each F1 in Majority Vote (MV), we report the baseline accuracy computed with the same item–annotator metric: for each item $x_i$, we collapse annotator labels $\{y_{ij}\}_j$ to a single consensus $y_i^{\mathrm{MV}} = \mathrm{mode}(\{y_{ij}\}_j)$ (ties broken deterministically by the training-set class prior), predict $y_i^{\mathrm{MV}}$ for all annotators, and score $\frac{|\{(i,j):\, y_i^{\mathrm{MV}}=y_{ij}\}|}{|\{(i,j)\}|}$.

### 6.3 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 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 3.

## 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.

**RINCE often 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. We further analyze per-annotator performance across methods on HSB in Appendix B.

**Multitask models degrade in performance with sparse annotators.** The multitask model performed the worst with the SNT dataset (Appendix A). This is likely due to how many annotators there are compared to how many samples they

| | 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 2: Mean Annotator F1 Scores (top row) with standard deviations (bottom row) showing the effect of $q$ on robustness across noise levels.

| Model | MDA | EPIC | HSB |
|---|---|---|---|
| Multitask | 0.498 | 0.642 | 0.806 |
| MichEmbed | 0.784 | 0.764 | **0.897** |
| AART | 0.795 | **0.790** | 0.891 |
| NRAEmbed | **0.796** | 0.789 | 0.891 |

Table 3: Mean Annotator F1 Scores comparing model robustness to renegade annotators (malicious/random annotation behavior). Bolded values highlight the best-performing approach across datasets (Results reported are average of 2 runs).

| 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 4: 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 samples, and #L is the number of labels.

annotated on average, which is very few, creating sparse annotators (see Table 4). This aligns with Mokhberian et al. (2023) , who showed through statistical parity analysis that multitask models exhibit the highest performance disparity for sparse annotators compared to embedding-based approaches, with more balanced and fair results. 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.

**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 A) show that contrastive methods underperform annotation-embedding approaches, suggesting ordinal settings require class-sensitive modifications.

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.

## 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 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 suggest 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.

## 9 Limitations

Our evaluation relies on synthetic noise injection (label flipping, random perturbations), 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 crowdsourced 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.

## 10 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.

In addition, we utilized the assistance of AI (LLMs) in some research, coding, and writing.

## 11 Acknowledgements

## References


Multilingual-Perspectivist-NLU/EPIC (epicorpus) dataset card. Hugging Face Datasets. Accessed 2026-01-04.

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.

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.

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 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. *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.

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*.

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

| Dataset | Noise | MV | Multitask | MichEmbed | AART | Ours |
|---|---|---|---|---|---|---|
| SNT | 0% | 0.245 | 0.267±.007 | 0.517±.009 | **0.521**±.006 | 0.494 |
| | 10% | 0.242 | 0.268±.002 | 0.492±.002 | **0.500**±.008 | 0.489 |
| | 20% | 0.251 | 0.259±.008 | **0.483**±.012 | 0.475±.016 | 0.471 |
| | 30% | 0.235 | 0.248±.013 | **0.446**±.001 | 0.439±.010 | 0.441 |
| | 40% | 0.236 | 0.225±.000 | **0.418**±.001 | 0.403±.021 | 0.406 |

Table 5: Mean annotator-level F1 on the multiclass SNT dataset under varying noise levels. Best is **bold**; second best is underlined. Unlike binary datasets, contrastive methods (AART, NRA-Embed) do not consistently outperform MichEmbed. MV = Majority Vote baseline.

# A 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 4).

## A.1 Noise Injection for Multiclass Labels

For SNT, we applied symmetric noise injection following a similar procedure to the binary datasets. Each annotators 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.

## A.2 Main Results on SNT

## A.3 Effect of Parameter $q$ on SNT

| | 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 ±.003 | 0.495 ±.005 | **0.502** ±.005 | **0.471** ±.013 | 0.463 ±.020 | 0.464 ±.011 | 0.400 ±.012 | 0.396 ±.019 | **0.400** ±.017 |

Table 6: 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 0% and 40% noise.

## A.4 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 5). 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 4). In contrast, multitask performed well on HSB, which has only 6 annotators who each labeled approximately 952 samples. This pattern aligns with findings from Mokhberian et al. (2023), 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."

## B Per-Annotator Analysis

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 7 presents F1 scores for each annotator across all methods and noise conditions.

### B.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 above 0.95 with embedding-based methods, and *low-agreement annotators* (A4, A5, A6) who score below 0.93. Notably, A5 consistently exhibits the lowest F1 across all methods (0.79–0.89 at 0% noise), suggesting this annotator's labeling patterns diverge most from learnable representations. This heterogeneity reflects genuine differences in annotator behavior rather than random variation, as the ranking remains stable across methods.

### B.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.

### B.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.

### B.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).

## C 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 the RINCE's original paper.

### C.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 is defined as:

$$\mathcal{L}_{\text{InfoNCE}}(s) = -\log\left(\frac{e^{s^+}}{Z}\right) = -s^+ + \log Z \quad (4)$$

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

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q}(s) = -\frac{e^{qs^+}}{q} + \frac{\lambda}{q} Z^q \quad (5)$$

Note that as $q \to 0$, Eq. (5) asymptotically recovers the behavior of the InfoNCE loss (up to a scaling factor).

Table 7: 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 |

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

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial s^+} = -(1-p), \quad \text{where } p = \frac{e^{s^+}}{Z} \tag{6}$$

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 ($\approx 1$).

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. (5) with respect to $s^+$ yields:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{RINCE}}}{\partial s^+} &= -e^{qs^+} + \lambda Z^{q-1} e^{s^+} \\ &= -e^{qs^+}\left(1 - \lambda\left(\frac{e^{s^+}}{Z}\right)^{1-q}\right) \\ &= -\underbrace{e^{qs^+}}_{\text{Weighting Term}} \cdot \underbrace{(1-\lambda p^{1-q})}_{\text{Error Term}} \end{aligned} \tag{7}$$

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), $e^{qs^+} \to 0$.

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.

### C.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 ($q \to 0$, 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 ($q \to 1$, 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.

For subjective sentiment analysis, where ground truth is often a distribution rather than a single fact, the pure exploration of InfoNCE is dangerous. By selecting a moderate $q$ (e.g., $q = 0.5$), RINCE enforces a necessary degree of exploitation, ensuring the model learns from high-consensus data while

remaining robust to the inherent noise of subjective annotation.

### C.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 lower bound on the Wasserstein Distance Maximization (WDM). 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), 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.