Don't Blame the Data, Blame the Model: Understanding Noise and Bias When Learning from Subjective Annotations

Abhishek Anand[1], **Negar Mokhberian**[1,2], **Prathyusha Naresh Kumar**[1], **Anweasha Saha**[1]
Zihao He[1,2], **Ashwin Rao**[1,2], **Fred Morstatter**[2], **Kristina Lerman**[2]
[1]Department of Computer Science, University of Southern California
[2]Information Sciences Institute, University of Southern California
{anandabh, nmokhber, nareshku, anweasha, zihaoh, mohanrao}@usc.edu
{lerman, fredmors}@isi.edu

Abstract

Researchers have raised awareness about the harms of aggregating labels especially in subjective tasks that naturally contain disagreements among human annotators. In this work we show that models that are only provided aggregated labels show low confidence on high-disagreement data instances. While previous studies consider such instances as mislabeled, we argue that the reason the high-disagreement text instances have been *hard-to-learn* is that the conventional aggregated models underperform in extracting useful signals from subjective tasks. Inspired by recent studies demonstrating the effectiveness of learning from raw annotations, we investigate classifying using Multiple Ground Truth (Multi-GT) approaches. Our experiments show an improvement of confidence for the high-disagreement instances[1].

1 Introduction

[Warning: This paper may contain offensive content.]

Datasets labeled by human annotators play a critical role in many supervised Natural Language Processing (NLP) tasks (Paullada et al., 2021). However, as the volume of such data has grown, it has become difficult to manually assess data quality. Recognizing this challenge, recent efforts have proposed automated strategies for evaluating annotated datasets, specifically targeting the identification of noisy and hard-to-learn data instances (Swayamdipta et al., 2020).

Existing methods for automatically gauging sample quality often rely on aggregated labels, such as a majority vote (Swayamdipta et al., 2020), but disagreements among annotations for data items are widespread (Plank, 2022). Some of these discrepancies arise from human labeling errors (Mokhberian et al., 2022), however, a growing body of research highlights that annotator differences in subjective tasks introduce bias in annotations, particularly in sensitive domains like hate speech recognition (Plank et al., 2014; Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Sap et al., 2022). Therefore, a single ground truth for each data instance may lead to potential oversights in capturing nuanced perspectives from different annotators.

In this paper, we leverage Data Maps (Swayamdipta et al., 2020), an automated data evaluation strategy, to understand the relation between noise and bias in annotated datasets. Data Maps define two intuitive measures for each data item: the model's confidence in predicting the true class and the variability of this confidence across epochs. Swayamdipta et al. (2020) have shown that lower model confidences correlate with higher chances of mislabeling for corresponding samples. Firstly, based on the assumption that a single correct label exists for a given example, we investigate an initial research question:

RQ1: *Is there any correlation between human disagreement on instances and model's uncertainty/confidence for classifying the instance to aggregated ground truth?*

Swayamdipta et al. (2020) has briefly studied the relationship between intrinsic uncertainty and the training dynamic measures. Their findings reveal a correlation between human disagreement and the model's uncertainty in a natural language inference dataset. We explore this correlation in the context of toxicity detection in social media texts using three different datasets. Our findings reveal a significant correlation between human label agreement and model confidence, with confidence decreasing as disagreements among annotators increase. Specifically, **s**ingle **g**round **t**ruth (Single-GT) models (see §4.1.2 for details) exhibit lower confidence for high-disagreement samples, potentially due to the subjectivity of those instances. These observations from *RQ1* motivate the exploration of **m**ultiple **g**round **t**ruths (Multi-GT) or

[1] Our code and data are publicly available at GitHub.

Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024), pages 102–113
March 22, 2024 ©2024 Association for Computational Linguistics

Multi-GT models (details in §5.1) that can infer based on multiple perspectives (Mostafazadeh Davani et al., 2022; Gordon et al., 2022; Weerasooriya et al., 2023; Mokhberian et al., 2023) as an alternative to Single-GT models.

As far as we are aware, there is limited existing research that has examined the training dynamics of non-aggregated annotations. Therefore, we adapt the Data Maps definition to Multi-GT models and empirically address our second research question:

RQ2:. *Does learning from raw annotations enhance the model's confidence for the high-disagreement instances?*

When using Multi-GT models, we identify improved confidence among minority votes for samples characterized by substantial annotation disagreements.

Our analysis in this paper demonstrates that samples receiving low confidence in Single-GT models are not inherently unusable. Furthermore, employing Multi-GT models for subjective tasks yields improved confidence for certain raw annotations associated with high-disagreement samples.

2 Related Work

Uncertainty in Machine Learning In the realm of uncertainty estimation and dataset evaluation, several studies have paved the way for understanding the dynamics of model training. Srivastava et al. (2014) introduce dropout-based uncertainty estimates, showcasing a positive relationship between training dynamics and dropout measures. The Data Maps approach (Swayamdipta et al., 2020) leverages this knowledge to establish the credibility of the proposed training dynamics measures and their relationship with uncertainty. Other works (Lakshminarayanan et al., 2017; Gustafsson et al., 2020; Ovadia et al., 2019) collectively support the notion that deep ensembles provide well-calibrated uncertainty estimates, laying the groundwork for our exploration of training dynamics measures and their correlation with uncertainty. Fort et al. (2020) sheds light on diversity trade-offs in ensembles, offering insights into the cost-effectiveness of using ensembles of training checkpoints. Chen et al. (2017) advocates for ensembles of training checkpoints as a more economical alternative with certain advantages. The work by (Xing et al., 2018) on loss landscapes provides additional perspectives on the optimization process during training, complementing the understanding gained from training dynamics.

Toneva et al. (2019) and (Pan et al., 2020) , along with (Krymolowski, 2002) , address catastrophic forgetting, providing approaches to analyze data instances. Bras et al. (2020) introduces AFLite, an adversarial filtering algorithm, advocating for the removal of "easy" instances. Chang et al. (2018) proposes active bias for training more accurate neural networks, aligning with the broader discussion on active learning methods presented in (Peris and Casacuberta, 2018; P.V.S and Meyer, 2019). Maddox et al. (2019) propose a technique for representing uncertainty in deep learning models utilizing Stochastic Weight Averaging to track a weighted average of neural network weights. Mishra et al. (2020) explores creating better datasets, resonating with the theme of dataset enhancement in the context of active learning methods. Influence functions (Koh and Liang, 2020), forgetting events (Toneva et al., 2019), cross-validation (Chen et al., 2019), Shapley values (Ghorbani and Zou, 2019), and the area-under-margin metric (Pleiss et al., 2020) contribute to the discussion on data error detection and instance scoring.

Multiple Perspectives In the paper by (Plank, 2022), the challenge of human label variation due to annotator perspective biases is described, emphasizing the impact on data quality, modeling, and evaluation stages. This resonates with our exploration of model confidence and the drawbacks of aggregating labels in subjective tasks. The call for Multi-GT designs aligns with our goal of understanding noise and bias in raw annotations.

The survey on 'Handling Bias in Toxic Speech Detection' by (Garg et al., 2023) provides insights into mitigating bias in toxic speech detection, reflecting the awareness raised by researchers about the harms of aggregating labels, especially in tasks involving disagreements among human annotators. This survey contributes relevant perspectives for enhancing the robustness and fairness of models in the context of subjective tasks.

Prior research has introduced models aimed at directly learning from annotation disagreements in subjective tasks. Two primary approaches have been proposed in this regard. The first approach treats the "ground truth" as the distribution encompassing all labels that a population of annotators could generate, as demonstrated in (Peterson et al., 2019; Uma et al., 2020). The second approach involves learning from the hard labels assigned by individual annotators, as explored in (Mostafazadeh Davani et al., 2022; Weerasooriya et al., 2023; Mokhberian et al., 2023).

While preceding studies have made significant strides in uncertainty estimation and dataset evaluation, our work adopts a novel perspective by questioning the effectiveness of aggregated models in identifying mislabeled samples. The definition of confidence used in this study and the Data Maps approach deviates from conventional usage in other fields, where confidence is typically assessed based on the predicted label. Alternative definitions and interpretations of confidence are present in certain core machine learning papers. The shift toward Multi-GT approaches and the exploration of diverse perspectives contribute to a more nuanced understanding of noise and bias within annotated datasets. The work by Wang and Plank (2023) suggests innovative uncertainty measures derived from Multi-GT models for integration into an Active Learning pipeline, aiming to decrease the budget required for item-annotator labeling. In contrast, our approach diverges as we focus on exploring training dynamics to capture noise in Multi-GT models.

3 Datasets

In this section we introduce the three datasets studied in this paper. Statistics of the datasets are presented in Table 1.

	$\mathcal{D}_{SI}$	$\mathcal{D}_{MHS}$	$\mathcal{D}_{MDA}$
# unique texts	45,318	39,565	10,440
# labels	2	3	2
# annotators	307	7,912	819
# annotations per text	3.2±1.2	2.3±1.0	5
# annotations per annotator	479±830	17±4	64±139

Table 1: The statistics for dataests introduced in §3

The social bias inference corpus ($\mathcal{D}_{SI}$) contains 45K posts from online social platforms such as Reddit, Twitter, and hate sites (Sap et al., 2020). The dataset includes structured annotations of social media posts with respect to offensiveness, intent to offend, lewdness, group implications, targeted group, implied statement, and in-group language. Following Weerasooriya et al. (2023) we only consider the labels from "intent to offend" for each data item.

The measuring hate speech corpus ($\mathcal{D}_{MHS}$) consists of 39,565 social media posts spanning YouTube, Reddit, and Twitter, manually annotated by 7,912 Amazon Mechanical Turk annotators from United States (Kennedy et al., 2020; Sachdeva et al., 2022). Annotations for each text sample include evaluating the intensity of 10 distinct hate speech labels, encompassing sentiment, disrespect, insult, humiliation, inferior status, violence, dehumanization, genocide, attack or defense, and hate speech. The labels are aggregated across all annotations for a given text using Rasch measurement theory (Rasch, 1960), resulting in a continuous hate speech score, where higher values denote increased offensiveness. This score is discretized into three labels: above +0.5 for hate speech, below -1.0 for supportive speech, and between -1.0 and +0.5 for neutral or ambiguous speech. We use these aggregated labels for Single-GT model. Furthermore, we incorporate each individual annotator's hate speech label as their specific annotation for Multi-GT model. Both the aggregated and non-aggregated target columns represent a multi-class classification task with 3 labels - supportive, neutral, or hate speech.

The Multi-Domain Agreement dataset ($\mathcal{D}_{MDA}$) has been created for studying offensive language detection (Leonardelli et al., 2021). It comprises approximately 400K English tweets from three topics: Covid-19, US Presidential elections, and the Black Lives Matter movement. Each tweet has been annotated for being offensive or not by 5 US native speakers using Amazon Mechanical Turk, resulting in a total of 10,753 annotated tweets. The tweets have been analysed further in Leonardelli et al. (2021) regarding level of annotator agreement: unanimous, mild, and low.

4 RQ1: Is there correlation between human disagreement and model's uncertainty/confidence?

4.1 Methods

This section outlines the approaches employed to address *RQ1*. We compute the agreement level in the human labels directly based on the annotations available in each dataset, with detailed explanations provided in §4.1.1. Subsequently, we investigate whether the classifiers' confidence in data items correlates with the level of annotator agreement.

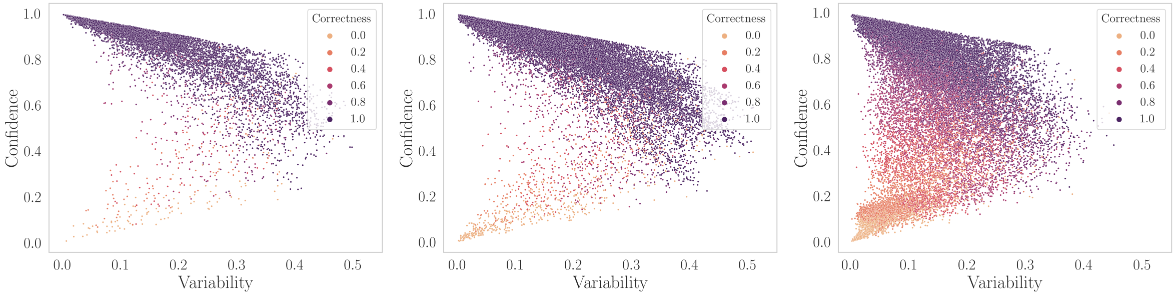

Figure 1: Dataset Cartography map for Single-GT model on $\mathcal{D}_{MDA}$ (left), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (right). The x-axis shows variability and y-axis, the confidence. Further, the points are color-graded by correctness (probability the trained model assigns this data point to the ground truth label in its prediction). Samples in the top left corner with high confidence and low variability are easy for the model to learn, whereas sample that are in the lower left corner with low confidence and low variability are difficult.

We utilize a conventional supervised text classification model, as elucidated in §4.1.2, and examine the training dynamics during defined epochs, outlined in §4.1.3.

4.1.1 Annotator Agreement Level

The annotator agreement level is defined as the proportion of annotations that align with the majority vote for a specific text sample. This metric, introduced by (Wan et al., 2023), provides insights into the degree of consensus among annotators regarding the majority label assigned to a given sample.

4.1.2 Single-GT Models

The conventional text classification model predicts the aggregated label for each instance. Text embeddings from transformer-based encoders are fed into a feed-forward classification layer which performs a linear projection layer to predict the majority label.

4.1.3 Data Maps

We adhere to the definitions outlined in Swayamdipta et al. (2020) to quantify the qualities of data instances automated by training classification models.

Confidence is defined as the mean class probability for each data item's gold label across all epochs. The confidence is tied to the evolution of class probabilities during the training process, offering insights into the model's certainty or consistency in predicting gold labels for each data item.

Variability is defined as the standard deviation of class probability for each data item's gold label across all epochs and measures the extent to which they change across different training epochs. It indicates the degree of fluctuation or stability in the model's predictions over time.

Swayamdipta et al. (2020) find that the simultaneous occurrence of low confidence and low variability correlates well with an item having an incorrect label.

4.2 Results

We calculate the training dynamics, confidence and variability to generate data cartography maps for the three datasets – $\mathcal{D}_{MDA}$, $\mathcal{D}_{SI}$ and $\mathcal{D}_{MHS}$ – as illustrated in Figure 1. Furthermore, we leverage training dynamics to evaluate the correlation between the model's confidence in predicting the gold label and the level of agreement among annotators for the gold label. This correlation is visually represented through boxplots in Figure 2 for Single-GT model where gold label is the aggregated vote. Across all three datasets, we identify a robust correlation between model confidence and annotator agreement level. Notably, instances of higher disagreement among human annotators correspond to lower model confidence throughout training epochs when the model is trained on the majority vote. To quantify the observed correlation, we utilize Pearson correlation coefficient with the results shown in Table 2 where we see large correlation for all three datasets with the associated p-values being statistically significant.

It is worth noting that the model remains unaware of annotator agreement level information during training, as it is only trained on a single ground truth label for a text sample, which is the majority vote. Nevertheless, this external factor significantly influences the model's confidence, with instances of heightened disagreement among human annotators corresponding to a persistent trend

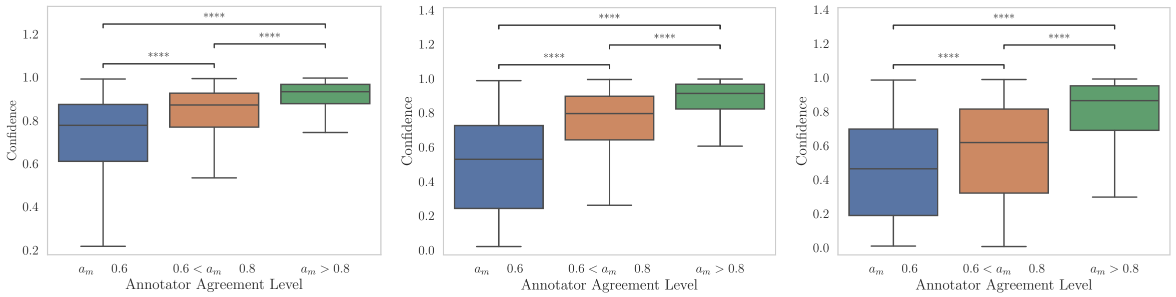

Figure 2: Boxplots illustrating the relationship between model confidence and annotator agreement level (a_m) for Single-GT model trained on $\mathcal{D}_{MDA}$ (left), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (right). There is a clear correlation between model's confidence in predicting the ground truth label and the agreement between annotators (denoted as the fraction of annotators that agree on the majority vote on the x-axis). We further depict significant differences in confidence distribution across agreement levels using the Mann-Whitney-Wilcoxon test (McKnight and Najab, 2010) with Statannotations (Charlier et al., 2022). Notation includes **** for $p <= 1.00e-04$.

Dataset	$\mathcal{D}_{MDA}$	$\mathcal{D}_{SI}$	$\mathcal{D}_{MHS}$
Corr.	0.44	0.37	0.45

Table 2: Pearson correlation coefficients between model confidence on each sample and the corresponding annotator agreement level for Single-GT model trained on the three datasets. The reported values are statistically significant.

of lower model confidence. Hence, a critical question arises: given the observed challenge where the model struggles to learn samples with high disagreement level exhibiting low confidence, can being exposed to multiple annotators' annotations enhance the model's learning capabilities on low confidence *(hard-to-learn)* samples?

5 RQ2: Do Multi-GT models lead to better confidences on hard-to-learn samples?

5.1 Methods

For our Multi-GT model, we rely on DisCo (Distribution from Context), as introduced by Weerasooriya et al. (2023), which is a neural model specifically designed for predicting labels assigned by individual annotators. Instead of considering items in isolation, this model takes annotator-item pairs as input and conducts inference by considering predictions from all annotators. The authors discover that incorporating annotator-specific modules into a classifier, as opposed to overlooking individual perspectives, leads to superior performance.

Following the DisCo model, in this study, the inputs consist of instance-annotation pairs $(x_m, y_{n,m})$, where x_m represents the mth data item, and $y_{n,m}$ denotes the label annotator n assigned to it. We adapt the calculation of confidence and variability based on the probabilities of gold annotation per instance-annotation. This approach yields multiple confidences per item, corresponding to the number of annotations available for that item.

Dataset	$\mathcal{D}_{MDA}$	$\mathcal{D}_{SI}$	$\mathcal{D}_{MHS}$
Corr.	0.46	0.44	0.51

Table 3: Pearson correlation coefficients between model confidence on each sample and the corresponding annotator agreement level for DisCo trained on the three datasets. When computing the training dynamics for DisCo, the pair of text sample and annotator ID is distinct across the dataset, which results in multiple confidence values for each annotation for a text sample. The reported values are statistically significant.

5.2 Results

As shown in the previous section, we employ training dynamics to assess the relationship between model confidence on annotations and the agreement level among annotators for a given text sample. We depict the relationship using Pearson correlation coefficient values in Table 3 with statistically significant p-values and the boxplots are illustrated in Appendinx A. It is important to note that for computing training dynamics for DisCo, the pair of text sample and annotator ID is unique across the dataset, hence, a text sample has multiple confidence values, one for each annotation for a text sample. We observe that consistent with the trend in models trained on a single ground truth label, heightened disagreement among annotators

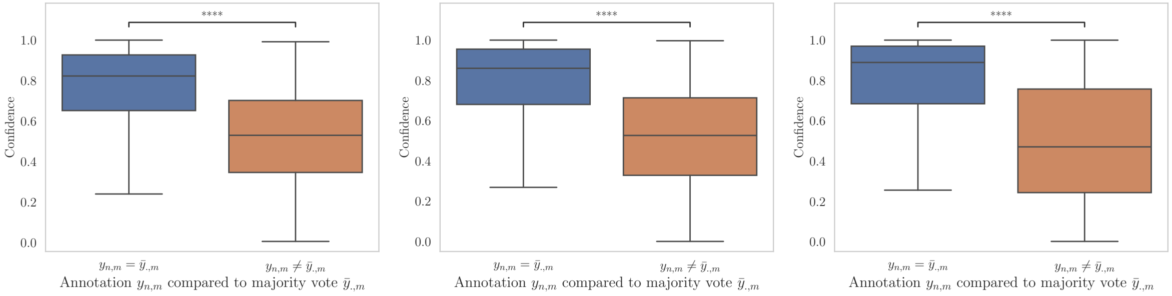

Figure 3: Boxplot illustrating the relationship between model confidence and whether the annotator's annotation ($y_{n,m}$) disagrees with the majority vote ($\bar{y}_{.,m}$) for DisCo trained on $\mathcal{D}_{MDA}$ (left), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (right). We see a clear correlation indicating higher confidence in the predicted label by the model when $y_{n,m} = \bar{y}_{.,m}$ and lower confidence when $y_{n,m} \neq \bar{y}_{.,m}$. We further depict significant differences in confidence distribution for $y_{n,m} = \bar{y}_{.,m}$ and $y_{n,m} \neq \bar{y}_{.,m}$ using the Mann-Whitney-Wilcoxon test (McKnight and Najab, 2010) with Statannotations (Charlier et al., 2022). Notation includes **** for $p <= 1e-04$.

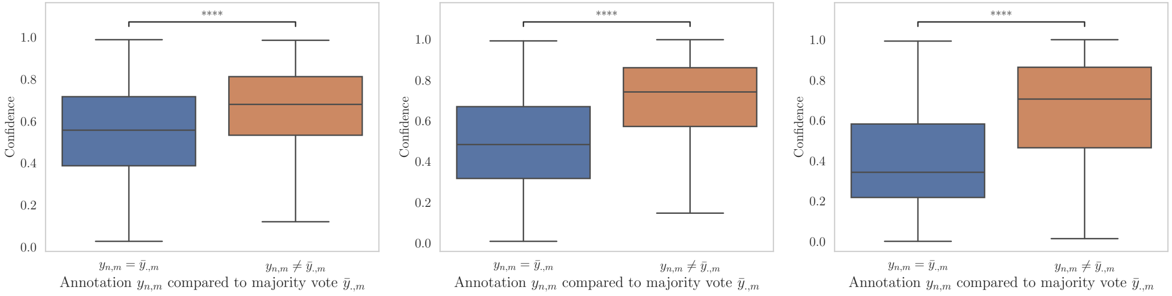

Figure 4: Boxplots illustrating the relationship between model confidence and whether the annotator's annotation ($y_{n,m}$) disagrees with the majority vote ($\bar{y}_{.,m}$) for DisCo trained on $\mathcal{D}_{MDA}$ (left), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (right) for DisCo only for the subset of samples where confidence is below 0.5 in Single-GT model. In contrast to the overall dataset presented in Figure 3, a reversed trend is observed, indicating higher confidence when $y_{n,m} \neq \bar{y}_{.,m}$ and lower confidence when $y_{n,m} = \bar{y}_{.,m}$. This highlights DisCo's ability to crucially learn from minority votes that are discarded for Single-GT model. We further depict significant differences in confidence distribution for $y_{n,m} = \bar{y}_{.,m}$ and $y_{n,m} \neq \bar{y}_{.,m}$ using the Mann-Whitney-Wilcoxon test (McKnight and Najab, 2010) with Statannotations (Charlier et al., 2022). Notation includes **** for p <= 1.00e-04.

for a text sample correlates with reduced model confidence. We further check the model confidence distribution for annotations ($y_{n,m}$), grouped by whether they are equal to majority vote ($\bar{y}_{.,m}$) for all three datasets depicted in Figure 3. The results show a clear trend: samples with $y_{n,m} = \bar{y}_{.,m}$ yield a high-confidence distribution, while those with $y_{n,m} \neq \bar{y}_{.,m}$ result in a notably lower confidence distribution. Two factors may contribute to this observation: 1) the inclusion of noisy minority vote annotations, where the majority vote represents an objectively correct label; and 2) the architectural limitations of the model. Although the model is designed to learn multiple annotations for a given text sample depending on the annotator ID as input, it encounters challenges in confidently learning the minority vote annotation for the text. These results emphasize the significance of annotator agreement in understanding uncertainty in model predictions, which applies to both Single-GT model and DisCo, a Multi-GT model, where higher confidence aligns with increased agreement on annotations.

Additionally, to answer the question whether DisCo, a Multi-GT model, is able to demonstrate increased confidence levels in hard-to-learn instances for the Single-GT model, our investigation specifically targets text samples where Single-GT model exhibits low confidence (below 0.5). As illustrated in Figure 4, a significant and consistent trend is observed across all three datasets. In this instance, samples with $y_{n,m} \neq \bar{y}_{.,m}$ show higher confidence compared to samples with $y_{n,m} = \bar{y}_{.,m}$. This contrasts with the relationship observed in the complete dataset boxplots in Figure 3, where model has higher confidence on samples with $y_{n,m} = \bar{y}_{.,m}$. This finding emphasizes a critical characteristic of DisCo, which can extract valuable information from annotations that are disregarded during the majority vote aggregation process. The Single-GT model never encounters this information and therefore cannot improve on challenging samples where the discarded annotation may be crucial due to mislabeled samples (Swayamdipta et al., 2020) or the subjectivity of the text.

We present a subset of the above group of samples with $y_{n,m} \neq \bar{y}_{.,m}$ in Table 4 that have high confidence in DisCo (above 0.9, i.e. easy to learn) and low confidence in Single-GT model (below 0.5) for $\bar{y}_{.,m}$. Provided with the opportunity to learn the minority vote label $y_{n,m}$ for these samples, DisCo rather finds it easy to learn them and hence, leading to the conjecture that majority votes $\bar{y}_{.,m}$ are inaccurate. We provide an additional set of examples in Appendix A where the Single-GT model exhibits high confidence (above 0.5) for $\bar{y}_{.,m}$, while DisCo demonstrates extremely low confidence (below 0.1) for $y_{n,m}$ where $y_{n,m} \neq \bar{y}_{.,m}$. This observation suggests that, in these instances, minority votes yn, m are deemed inaccurate.

Further, to evaluate the model's capability to learn multiple annotator perspectives, we focus on samples with disagreement in the dataset where annotator agreement level is below 1.0, signifying disparate labels provided by different annotators for the same text. Effectively capturing diverse annotator perspectives entails the model's ability to accurately predict distinct labels for identical text inputs based on annotator input, showcasing its ability to learn varied perspectives encoded in the annotations. To illustrate this, in Figure 5 we plot the count of samples with disagreement grouped by the number of different labels the model learns with high confidence (above 0.5). This visualization would help us assess whether the model is able to learn multiple labels for a text with high confidence, when the sole variation in input to the model lies in the annotator ID. Thus, it serves as an evaluation of its capability to learn different annotator viewpoints. For datasets $\mathcal{D}_{MDA}$ and $\mathcal{D}_{SI}$, with a binary classification task, although the model confidently learns a single label for over 50% of the samples, there is still a notable subset of samples (All Labels > 0.5), where the model shows high confidence for both labels, indicating its ability to capture annotator perspectives.

However, for $\mathcal{D}_{MHS}$, characterized by three labels, insights from Figure 5 reveal that DisCo confidently learns only a single label for over 75% of the samples, with approximately only 12% samples where it confidently learns multiple labels. This underscores its challenge in capturing individual annotators' perspectives through their annotations. We attribute this difficulty to the notably low average number of annotations per annotator in $\mathcal{D}_{MHS}$ (below 20), as shown in Table 1, in contrast to the other two datasets. The limited number of annotations per annotator presents an obstacle in effectively modeling an annotator's perspective. Therefore, we emphasize that accumulating a substantial number of annotations from each annotator is imperative for the effectiveness of DisCo.

Our analysis unveils key insights into model confidence and annotation dynamics. Examining the relationship between model confidence and annotator agreement levels for text samples, our findings echo those in Single-GT models, showing that heightened annotator disagreement aligns with decreased model confidence. In hard-to-learn instances for the Single-GT model, DisCo showcases increased confidence in samples with minority vote annotations, revealing its capacity to extract valuable insights from annotations typically overlooked in majority vote aggregation. Moreover, our investigation reveals that DisCo can effectively predict diverse labels for identical text inputs, especially in instances marked by disagreement, but it struggles in datasets with a limited number of annotations per annotator, emphasizing the necessity of accumulating a substantial number of annotations for DisCo's effectiveness. In essence, our findings underscore the critical importance of preserving multiple perspectives through annotations in subjective tasks and advocate for advancements in modeling approaches to achieve nuanced learning for broader representations.

6 Conclusions

This paper delves into an exploration of whether perspectivist classification models effectively harness valuable insights from instances identified as noisy through automated dataset evaluation techniques. Our investigation begins by examining how Single-GT models classify high-disagreement elements as noise. Subsequently, we shift our approach to Multi-GT models and observe a notable increase in confidence for minority votes for the same instances. This shift underscores the potential for richer and more nuanced understanding when leveraging multiple perspectives in the classification process.

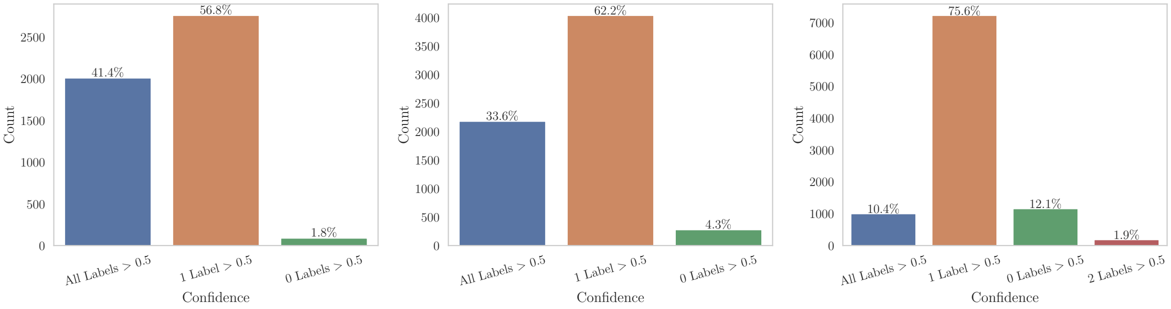

Figure 5: Number of samples with disagreement in the dataset where annotator agreement level is strictly below 1.0 and they are grouped by the number of different labels DisCo learns with high confidence (above 0.5) for these samples for the dataset $\mathcal{D}_{MDA}$ (left), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (right).

Dataset	Text	Majority Label	Minority Label
$\mathcal{D}_{MDA}$	START PUNCHING BACK !!!	offensive	not offensive
	Y'ALL FORGOT THE WHITE DUDE THAT THE POLICE KILLED LAYING JUST LIKE GEORGE FLOYD THE CRIMINAL <url>	not offensive	offensive
	His life mattered. You however say whatever sounds good. It's disgusting. You insult this issue.	offensive	not offensive
$\mathcal{D}_{SI}$	why are jews ' noses so big ? air is free alt : more gold to dig	not offensive	offensive
	What did the Texas shooter say as he walked through the church? Pew pew pew!	not offensive	offensive
	Still better that Nero's poetry. // cc @ChrisWarcraft	offensive	not offensive
$\mathcal{D}_{MHS}$	Fucking fagot burn in hell	supportive	offensive
	Your nigga don't suck toes or eat ass he a bitch.	supportive	offensive
	At 7:19 is why I hate people, women especially look how mad she got just because the dog splashed her.. f*** you you stupid b**** either have fun or go kill yourself	neutral	offensive

Table 4: Examples from the three datasets $\mathcal{D}_{MDA}$, $\mathcal{D}_{SI}$ and $\mathcal{D}_{MHS}$ where Single-GT model has low confidence (below 0.5) for the Majority Label and DisCo has really high confidence (above 0.9) for the Minority Label. Following our best assessment, it appears that the majority label for this subset appears to be inaccurate, and the minority label emerges as the more suitable annotation.

For future research directions, it is worth exploring model confidences for each annotator in the dataset in the context of the Multi-GT model. This investigation will enhance our understanding of the challenges faced by current models in learning annotator perspectives. Additionally, it is also worth exploring datasets like $\mathcal{D}_{MHS}$ featuring annotator demographic details and target demographic information for offensive text. Such datasets provide a chance to assess model confidences for both Single-GT and Multi-GT models across diverse demographic groups. This presents an opportunity to investigate the impact of preserving diverse perspectives through annotations in addressing societal biases within learned models.

Limitations

Although we have carried out a comprehensive analysis, our study has certain limitations that warrant consideration. Firstly, the performance of Multi-GT models is dependent on the number of annotations per annotator, and a low number in some datasets may impact the representation of individual annotators. Secondly, the absence of raw annotations in many datasets limits a broader analysis of potential bias or noise. Additionally, variations in annotation instructions across datasets and differing levels of freedom for subjective interpretation among annotators introduce potential biases and inconsistencies that may affect comparison. Moreover, for Multi-GT models, this paper only considers DisCo, which requires an annotator ID to make the prediction. However, future research can explore the models that learn from the distribution of labels for each item. Furthermore, various approaches to defining annotators' label agreement, such as entropy and silhouette score (Mokhberian et al., 2022), could be explored in forthcoming research. Finally, despite employing a Multi-GT approach, there is a possibility that the dataset items and annotators may have limitations as they may belong to a non-representative pool that does not encompass diverse societal perspectives. These limitations highlight the importance of cautious interpretation and generalization of our findings.

Ethical Considerations

We employ Multi-GT models to capture diverse perspectives in the classifier. However, it's conceivable that the items or annotators within each collected dataset may be constrained in various ways, and the annotator pool may not accurately represent perspectives from the entire societal spectrum. Limitations could stem from factors such as an insufficient count of annotators from specific demographics in the pool or the presence of noisy annotations from certain annotators.

An additional ethical consideration in training Multi-GT models that capture the preferences of individual annotators is the issue of privacy and anonymity. It is crucial to ensure that annotators remain anonymized, and the process of learning and inferring their personal perspectives is conducted in a manner that avoids any potential misuse or harm.

Acknowledgments

This work was funded in part by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002. We express gratitude to the anonymous reviewers for providing valuable feedback and offering suggestions for our project.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2018. Active bias: Training more accurate neural networks by emphasizing high variance samples.

Florian Charlier, Marc Weber, Dariusz Izak, Emerson Harkin, Marcin Magnus, Joseph Lalli, Louison Fresnais, Matt Chan, Nikolay Markov, Oren Amsalem, Sebastian Proost, Agamemnon Krasoulis, getzze, and Stefan Repplinger. 2022. Statannotations.

Hugh Chen, Scott Lundberg, and Su-In Lee. 2017. Checkpoint ensembles: Ensemble methods from a single training process.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. 2019. Understanding and utilizing deep neural networks trained with noisy labels.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2020. Deep ensembles: A loss landscape perspective.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s).

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning.

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

Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. 2020. Evaluating scalable bayesian deep learning methods for robust computer vision.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Pang Wei Koh and Percy Liang. 2020. Understanding black-box predictions via influence functions.

Yuval Krymolowski. 2002. Distinguishing easy and hard instances. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles.

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.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: Measuring data quality in nlp.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 147–154. IEEE.

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

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.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. 2020. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems*, volume 33, pages 4453–4464. Curran Associates, Inc.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

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

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626.

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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking.

Avinesh P.V.S and Christian M. Meyer. 2019. Data-efficient neural text compression with interactive learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2543–2554, Minneapolis, Minnesota. Association for Computational Linguistics.

Georg Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. 2018. A walk with sgd.

A Supplemental Material

A.1 Experimental Setup

Dataset	F1 (Single-GT)	F1 (DisCo)
$\mathcal{D}_{MDA}$	0.78	0.78
$\mathcal{D}_{SI}$	0.80	0.78
$\mathcal{D}_{MHS}$	0.68	0.75

Table 5: F1(weighted) scores for Single-GT and DisCo trained on the three datasets.

For our experiments we utilize pre-trained RoBERTa-Base as Single-GT model for fine tuning on $\mathcal{D}_{MDA}$, $\mathcal{D}_{SI}$ and $\mathcal{D}_{MHS}$. Both Single-GT and DisCo were trained for 5 epochs on each dataset to compute training dynamics values of confidence and variability. We report the F1 scores for Single-GT and DisCo models trained on $\mathcal{D}_{MDA}$, $\mathcal{D}_{SI}$ and $\mathcal{D}_{MHS}$ in Table 5 offering a summary of their performance to highlight convergence of models and reliability of predictions.

A.2 Additional Examples and Plots

Dataset	Text	Majority Label	Minority Label
$\mathcal{D}_{MDA}$	Shocking! (Sorry, sarcasm doesn't work well on Twitter)	not offensive	offensive
	<user> Did he really die of COVID-19. ???	not offensive	offensive
	Fragility at its finest	not offensive	offensive
$\mathcal{D}_{SI}$	Rate this Anti-Social Socially Anxious Hermit Loser with Long Hair (Pic)	offensive	not offensive
	What's a pedophile's all you can eat buffet? The school bus	offensive	not offensive
	But is it really that big a deal?	not offensive	offensive
$\mathcal{D}_{MHS}$	@Cam_Coles You don't know what I know. Name somebody colder, you monkey faced immigrant	offensive	supportive
	When these niggas force me to say good morning URL	offensive	neutral
	Lailat al Miraj mubarak to all Muslims.. I'm wishing for everyone's peace & prosperity	supportive	neutral

Table 6: Examples from the three datasets $\mathcal{D}_{MDA}$, $\mathcal{D}_{SI}$ and $\mathcal{D}_{MHS}$ where Single-GT model has high confidence (above 0.5) for the Majority Label and DisCo has really low confidence (below 0.1) for the Minority Label. Following our best assessment, it appears that the minority label in this case appears to be inaccurate.

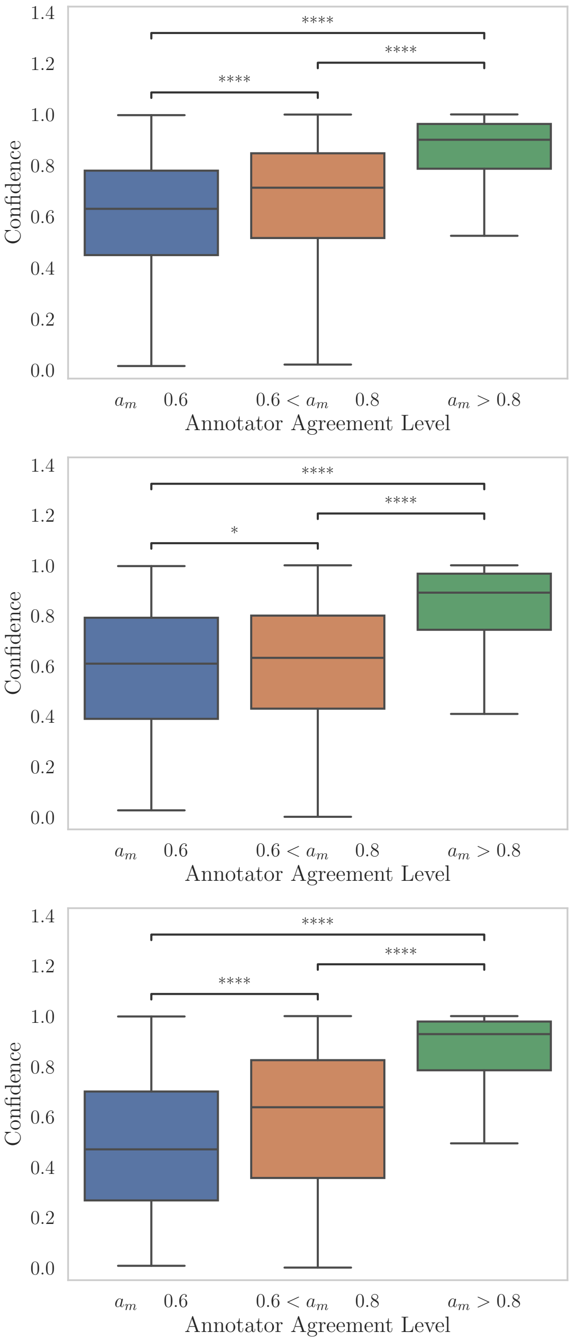

Figure 6: Boxplots illustrating the relationship between model confidence and annotator agreement level (a_m) for DisCo trained on $\mathcal{D}_{MDA}$ (top), $\mathcal{D}_{SI}$ (center) and $\mathcal{D}_{MHS}$ (bottom). We see a clear correlation indicating higher confidence in the predicted label by the model with higher agreement between annotators (denoted as the fraction of annotators that agree on the majority vote on the x-axis). We further depict significant differences in confidence distribution across agreement levels using the Mann-Whitney-Wilcoxon test (McKnight and Najab, 2010) with Statannotations (Charlier et al., 2022). Notation includes * for $1e-02 < p <= 5e-02$ and **** for $p <= 1e-04$.