Hypernetworks for Perspectivist Adaptation

Daniil Ignatev[1], Denis Paperno[1], Massimo Poesio[1,2],

[1]Utrecht University, [2]Queen Mary University of London,
Correspondence: d.ignatev@uu.nl

Abstract

The task of perspective-aware classification introduces a bottleneck in terms of parametric efficiency that did not get enough recognition in existing studies. In this article, we aim to address this issue by applying an existing architecture, the `hypernetwork+adapters` combination, to perspectivist classification. Ultimately, we arrive at a solution that can compete with specialized models in adopting user perspectives on hate speech and toxicity detection, while also making use of considerably fewer parameters. Our solution is architecture-agnostic and can be applied to a wide range of base models out of the box.

1 Introduction

In the recent years, perspective-aware approach to subjective linguistic tasks has been gaining prominence in NLP. This approach suggests that for tasks that involve subjectivity, dataset designers should collect multiple labels from different annotators for each data instance; these labels need to be retained and used in model training (Plank, 2022; Cabitza et al., 2023; Fleisig et al., 2024). This policy is warranted in tasks like hate speech detection, where multiple labels assigned by annotators with diverse backgrounds can be equally applicable. This notion contrasts the popular practice of designating a single supposedly correct label for each bit of data while discarding conflicting annotator judgments.

Likewise, NLP researchers have argued that subjective tasks require perspective-aware machine learning methods, i.e., methods that can capture diverse opinions based on unaggregated labels (Akhtar et al., 2021). We are particularly interested in the paradigm of *strong perspectivism* (Cabitza et al., 2023). The latter suggests that much like model personalization, separate models or representations should be trained on labels produced by individual annotators. In practice, this entails one of the following: (1) training a full language model checkpoint on each annotator's labels; (2) when using PEFT methods (Houlsby et al., 2019), training a separate adaptation component for each perspective; (3) using a specialized language model architecture.[1] We argue that both the first and the second strategy require a prohibitive amount of trainable parameters that increases with model size and number of modeling targets (see discussion in Section 6). To tackle this problem, we make use of an architecture that employs a small trainable module and adapts a model to diverse perspectives while keeping most parameters frozen.

The core of our solution is the `hypernetwork+adapters` combination. A hypernetwork is a neural architecture in which a source neural network is trained to predict the weights of a target network; these weights are subsequently used in inference (Ha et al., 2017). Importantly, hypernetworks can also be used to predict weights of adapters, including LoRA adapters (Houlsby et al., 2019; Hu et al., 2022), rather than weights of entire models. Adapters can be defined as small trainable modules designed for parameter-efficient tuning of NLP models including large language models; compatibility with adapters makes hypernetworks suitable for the same task (Karimi Mahabadi et al., 2021; He et al., 2022; Phang et al., 2023). We provide more details on this architecture in Section 3. The novel aspect of our work is that we attempt to repurpose the `hypernetwork+adapters` combination for perspectivist modeling and, by extension, model personalization. We posit its strengths, i.e., the reduction of trainable parameters, as a way to address the mentioned bottleneck issue. To our knowledge, this direction has not been thoroughly explored in previous studies.

Our article makes the following con-

[1]Here, we do not consider few-shot and zero-shot learning, as they do not fall under the training category.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 111–122

tributions. First, we consider whether `hypernetwork+adapters` setting is generally applicable to annotator-aware text classification. Our results suggest that hypernetworks are well fit for this task, albeit not unconditionally.

Second, in our experiments,[2] we show that our hypernetwork-based architecture performs on par with recent perspectivist model architectures — particularly, Annotator-Aware Representations for Texts (Mokhberian et al., 2024), and Annotator Embeddings (Deng et al., 2023). At the same time, we note the limitations of the proposed architecture and attempt to address them.

Third, we demonstrate that the proposed architecture offers a better trade-off in terms of parameter efficiency than the baseline models. At the same time, it also demonstrates a greater degree of versatility: since the weights of the base model are not being affected, there is no risk of forgetting (Kirkpatrick et al., 2017) or degradation of the base model's fundamental capabilities. These properties of our method make us believe that it has promise in modeling subjective tasks.

2 Related work

Data perspectivism: the idea of preserving pluralistic annotations in datasets has been discussed for a considerable time (Artstein and Poesio, 2008) and has been gaining increasing prominence across various AI domains (Kumar et al., 2021; Kapania et al., 2023; Huang et al., 2023). Frenda et al. (2024) provides a detailed survey of such perspectivist datasets and methods, reflecting a paradigm shift towards treating annotator disagreement not as noise but as a source of valuable signal about human variation.

Perspectivist learning: modeling human variation in labeling based on annotator perspectives has been advocated in several recent studies (Plank, 2022; Cabitza et al., 2023). In terms of modeling techniques, various recipes have been explored, including trainable crowd layers (Rodrigues and Pereira, 2018); trainable tokens (Sarumi et al., 2024); multi-task classification with annotator-specific heads (Davani et al., 2022), and others. Recently, studies have focused on active learning (Baumler et al., 2023; Wang and Plank, 2023) and few-shot perspective modeling (Golazizian et al., 2024; Sorensen et al., 2025) as two ways to deal with data sparsity.

Strong perspectivism brings together perspectivist learning and model personalization, which is why recent methods from the latter domain are also relevant to our research. In particular, studies by Tan et al. (2024) and Clarke et al. (2024) show that adapters and LoRA adapters can rival full model finetuning on the LLM personalization task, even when the number of trainable parameters is reduced to a fraction of the original model size.

Finally, in our study, we are particularly interested in recent works that optimize encoder-based classifiers against pluralistic labels and make use of annotator embeddings (Deng et al., 2023; Mokhberian et al., 2024). In this work, we consider both of these as our baselines.

Hypernetworks: The idea of hypernetworks has its roots in an earlier concept of weight generators (Gomez and Schmidhuber, 2005) and aims to constrain the search space when modeling complex objectives through searching in the limited weight space. Practically, this means generating the weights of a target model by means of a separate generator model (hypernetwork).

The title paper, published in 2017 (Ha et al., 2017), had two additional objectives: reduction of trainable parameters (1) and regularized training (2). The study proposed an implementation of two network types: a CNN network generating weights for another CNN network and an RNN network producing weights for a target RNN. Their results showed that the system achieves competitive performance and reduces the trainable parameter count, effectively fulfilling its purpose. The authors' assumptions were further scrutinized in follow-up works. As an example, a study by Soydaner (2020) applied hypernetworks to convolutional autoencoder models and showed that the weights of an autoencoder can be closely approximated by a much smaller hypernetwork, resulting in significant reduction of trainable parameters.

Moreover, the ML community recognized early the potential of hypernetworks in multitask learning. For instance, a study by Tay et al. (2021) adapts a transformer model to a multitude of tasks by predicting task-specific weights for the model's feed-forward layer using a hypernetwork. This adaptation strategy is of particular interest to us due to Davani et al. (2022)'s success in learning perspectives through multi-task classification.

[2]Our codebase has been made publicly accessible at https://github.com/ruthenian8/Hypernets

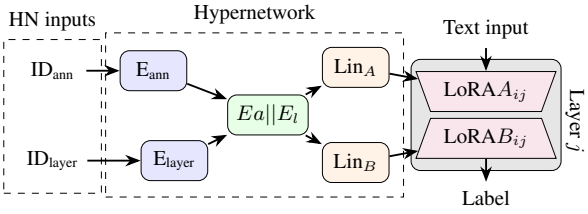

Figure 1: Data flow within our hypernetwork implementation: the annotator id and the layer id are embedded to adapt the target layer l_j in the base model.

The above architectures leveraged hypernetworks to predict the model's own weights. However, the parameter-efficient finetuning paradigm brought about an alternative approach, which is to infer the weights of small trainable submodules while leaving the weights of the main model intact. Adapters are a specific instance of these submodules (Houlsby et al., 2019); modeling them with hypernetworks has led to impactful results in several machine learning applications. Specifically, in image generation, hypernetworks have been used to personalize diffusion models achieving a considerable speed-up compared to other methods (Ruiz et al., 2023). In a different vein, in speech recognition, Müller-Eberstein et al. (2024) employed hypernetworks to adapt ASR models to individual speakers with atypical speech patterns. Their hypernetwork obtains a 75% relative reduction in word error rate using only 0.1% of the model parameters.

In NLP, hypernetworks have been utilized for multi-task and multilingual adaptation of larger transformer models, such as T5 (Raffel et al., 2020). Üstün et al. (2022) propose a single hypernetwork that produces adapter weights for multiple languages and tasks simultaneously, eliminating the need for separate language-specific task adapters. Karimi Mahabadi et al. (2021) and Phang et al. (2023) leverage a shared hypernetwork to effectively train adapter modules for a range of NLP tasks. Finally, in 2025, Charakorn et al. (2025) used hypernetworks to generate task-sepecific LoRA parameters for LLMs on the fly given a textual task description, thus enabling zero-shot adaptation; their results show that hypernetworks are combinable with LLMs, while their approach may also be repurposed for user personalization in the future.

3 Method

3.1 Architecture

The proposed architecture aims to tune a base model to each annotator's perspective; it implements an adapter-modelling hypernetwork in a fashion similar to Phang et al. 2023 and Müller-Eberstein et al. 2024, as it specifically makes use of low-rank adapters (Hu et al., 2022). Generally, all adapter variations enable parameter-efficient finetuning of large models by freezing and patching the base model's pre-trained weights W_j within a layer l_j with a smaller trainable layer Ada_j. Low-rank adapters, in particular, take this reduction of updatable parameters one step further by decomposing Ada_j into two low-rank matrices, A_j and B_j; ultimately, they decrease the parameter count even more with little performance impact. This principle can thus be illustrated with the following formula:

$$W'_j = W_j + B_j A_j \tag{1}$$

where A_j and B_j are low-rank trainable matrices. We find this type of adapters preferable for use in combination with hypernetworks, as it narrows down the solution space for the hypernetwork component.

In our implementation, when modeling the perspective Ann_i, we use the hypernetwork H to predict A_{ij} and B_{ij} for every layer l_j. Hence, we need to condition our hypernetwork on two relevant variables: information on the target annotator Ann_i and on the target layer l_j.

$$A_{ij}, B_{ij} = H(Ann_i, l_j) \tag{2}$$

For demonstration purposes, we only consider unique annotator IDs as annotator information. However, other relevant variables can also be straightforwardly integrated. We leave it to further research to determine whether using more complex representations based on sociodemographic variables or prior annotations shows better efficiency.

Likewise, target layer information is supplied through numeric layer identifiers. In the hypernetwork module, both types of identifiers are embedded, concatenated, and jointly passed to two prediction heads (Lin_A, Lin_B), which then infer the matrices A and B; this flow is illustrated in Figure 1.

We add the hypernetwork component on top of PEFT's architecture-agnostic LoRA implementation (Mangrulkar et al., 2022) with the aim of making our method compatible with a wide range of base models. To mitigate possible generalization issues, we use GeLU activation (Hendrycks and Gimpel, 2016) and dropout probability of 0.25 in the hypernetwork module. We also follow the suggestion of Müller-Eberstein et al. 2024 by initializing Lin_B with zeros and Lin_A with values close to zero, ensuring smooth updates at the initial stages of training.

3.2 Data

In this study, we purposefully pick 4 datasets that permit us to compare our architecture against existing algorithms for perspectivist learning. Two of these datasets, HS-Brexit and MD-Agreement, were included in the shared task on Learning With Disagreements (LeWiDi, Leonardelli et al. 2023). This section gives an overview of all data we used.

The Multi-Domain Agreement dataset ($\mathcal{D}_{MDA}$): This dataset by Leonardelli et al. (2021) addresses the task of offensive language detection. MD-Agreement comprises 9,814 English tweets from three distinct domains: the Black Lives Matter movement, the 2020 Election, and the COVID-19 pandemic. Each tweet was annotated by at least 5 Mechanical Turk workers out of a total pool of 334.

English Perspectivist Irony Corpus ($\mathcal{D}_{EPIC}$): Introduced by Frenda et al. (2023), the corpus consists of 3,000 Post-Reply pairs collected from Twitter and Reddit. The data was sourced from five English-speaking countries: Australia, India, Ireland, the United Kingdom, and the United States. A total of 74 annotators, balanced by gender and nationality, participated in the annotation task, with around 15 raters per nationality. Each annotator labeled approximately 200 instances, resulting in a corpus with 14,172 annotations and a median of 5 annotations per instance.

The Racial Bias Toxicity Detection Corpus ($\mathcal{D}_{RB}$): Sap et al. (2019) investigated the interaction between annotators' biases and their perceptions of toxicity reaching positive conclusions. The authors recruited 819 Amazon Mechanical Turk workers to annotate tweets for two variables: whether a tweet was (a) personally offensive to them and (b) potentially offensive to others. As in Mokhberian et al. 2024, our experiments focus on (a).

Hate Speech Brexit ($\mathcal{D}_{HS\text{-}Brexit}$): Akhtar et al. (2021) had 6 experiment participants label the same set of 1120 tweets related to Brexit. Each tweet was annotated for several categories including Hate Speech, Aggressiveness, Offensiveness, and Stereotype. In our study, we focus on Hate Speech annotations.

Dataset	*Train*	*Dev*	*Test*	*#A*	*#E/#A*
$\mathcal{D}_{MDA}$	$27k$	$13k$	$13k$	334	160
$\mathcal{D}_{EPIC}$	$7.1k$	$3.5k$	$3.5k$	74	191
$\mathcal{D}_{RB}$	$6.1k$	$2.7k$	$2.8k$	819	14
$\mathcal{D}_{HS\text{-}Brexit}$	$3.3k$	$1.6k$	$1.6k$	6	1120

Table 1: Post-split statistics of datasets used in the experiments. *#A* stands for the number of workers; *#E/#A* denotes the mean number of annotated items per worker.

For our experiments, we reproduce the dataset splitting procedure from Mokhberian et al. 2024. Specifically, we split the data into partitions of 50, 25, and 25% (train, dev, and test sets respectively) stratified with respect to item-level disagreements. Items from the dev and test sets annotated by an annotator not present in the training set are merged into the training data. This splitting procedure is repeated using 10 different random seeds. We report the post-preprocessing statistics for the four datasets in Table 1.

4 Experiments

4.1 Baselines

We test our architecture against approaches introduced in Deng et al. 2023 (AE) and Mokhberian et al. 2024 (AART). Both of these build on generic transformer models and handle diverse perspectives by integrating them directly into the architecture. To that end, they make use of specialized annotator representations.

AART: The AART architecture combines text embeddings from pretrained transformer models with learned annotator embeddings. Formally, given a text item x_i and annotator a_j, the combined embedding is computed as:

$$g(x_i, a_j) = e(x_i) + f(a_j)$$

where $e(x_i)$ is the text embedding, and $f(a_j)$ is a learned annotator-specific embedding. This combined embedding is then fed into a common

classification head. The model employs a multi-part loss function consisting of cross-entropy loss, L2 regularization on annotator embeddings, and a contrastive loss designed to cluster similar annotator perspectives. We compare our results against AART on $\mathcal{D}_{MDA}$, $\mathcal{D}_{EPIC}$, and $\mathcal{D}_{RB}$.

Annotator Embeddings (AE): This approach explicitly captures annotator-specific biases and annotation tendencies using two types of embeddings: annotator embeddings (E_a) and annotation embeddings (E_n). These embeddings are combined with the original text embeddings via weighted summation:

$$E_{\text{combined}} = E_{[\text{CLS}]} + \alpha_n E_n + \alpha_a E_a$$

where α_n and α_a are learnable weights computed based on the interaction between the sentence embedding and annotator-specific embeddings. The resulting embedding is fed into transformer-based classification models, enhancing their ability to predict annotator-specific labels by explicitly modeling annotator idiosyncrasies. We compare the performance of our model against AE on $\mathcal{D}_{MDA}$ and $\mathcal{D}_{HS\text{-}Brexit}$.

Single-task baseline: Both AART and AE compare their systems against a single-task baseline model. This baseline is a transformer classifier trained on majority-aggregated labels that predicts one label per text and effectively ignores annotator-specific nuances. This baseline mostly serves as a sanity check to evaluate the effectiveness of perspectivist modeling under the assumption that specialized architectures should handle label diversity at least somewhat better than a non-specialized model. However, on data items where most annotators agree with the majority label (e.g., 7 raters out of 10), it can show better results than a perspectivist model due to a much narrower problem space. Such data items constitute a large part of the existing perspectivist datasets, and, as a result of that, specialized models do not always surpass this baseline.

4.2 Setup

In all experiments, we use RoBERTa-base (Liu et al., 2019) as our base model to maintain consistency with the baseline methods, as both studies include reported results for RoBERTa-base among other models. As a part of our solution, all layers except the hypernetwork get explicitly frozen. We model LoRA adapters with a rank of 2 and $\alpha = 32$, keeping adapter dimensionality to a minimum. Although increasing the number of trainable parameters is likely to lead to better results, we explore the lower performance boundary of our method that also offers the best efficiency tradeoff.

We train the hypernetwork for 5 epochs with a batch size of 100 at a learning rate of $1\text{e}{-5}$ making use of Adam optimizer (Kingma and Ba, 2014). We set max. sequence length to 100. For each dataset, we repeat the experiment 10 times with varying random seeds. To implement our method, we use abstractions from PEFT (Mangrulkar et al., 2022) and Transformers (Wolf et al., 2020).

We obtained the above hyperparameter values by running a grid search with varying dropout probabilities (0.0, 0.1, 0.25) and learning rates ($5\text{e}{-5}$, $1\text{e}{-5}$, $5\text{e}{-6}$) using $\mathcal{D}_{EPIC}$'s development set as a reference. Our conclusion is that unlike the baseline solutions without PEFT methods, our architecture shows greater sensitivity to learning rate and dropout probability. In particular, when using low dropout (0 or 0.1) or a higher learning rate ($5\text{e}{-5}$), the model often converges prematurely at a suboptimal point. This yields micro-f1 values of ≈ 50.0 suggesting that the model only learns the majority label due to label imbalance; in contrast, when using our ultimate parameter values (dropout$= 0.25$, lr$= 1\text{e}{-5}$), training proceeds with gradual parameter updates and leads to better generalization (≈ 68.5). This result demonstrates that careful hyperparameter tuning is necessary for our approach.

4.3 Evaluation

Since AART and AE were assessed with two different metric suites, we preserve the original evaluation protocols to allow direct comparison with the published figures. In addition, we track the total number of trainable parameters for each model as described below.

Annotator-level F1 This metric measures the ability of a model to treat every annotator fairly, regardless of how many labels they contributed. For each annotator a_j, we compute the macro-F1 score over all items x_i in the test split by comparing the gold label y_{ij} to the model's prediction $\hat{y}_{ij}$. The *Annotator-level F1* is then the simple mean of these per-annotator F1 scores. Mokhberian et al. 2024 argue that it prevents evaluation biases towards prolific annotators.

Global-level F1 To measure overall predictive quality, we pool all (x_i, a_j) pairs in the test set and compute macro-F1 on this combined set. Unlike the annotator-level metric, this score weighs each prediction equally and inadvertently prioritizes annotators who contributed more labels.

Item-level Disagreement Correlation We quantify how well a model reproduces the true pattern of annotator disagreement on each item. For an item x_i with annotator votes $\{y_{i1}, \ldots, y_{iK}\}$, the gold disagreement is

$$d_i = 1 - \frac{\max_c \left|\{j : y_{ij} = c\}\right|}{K},$$

and the model-predicted disagreement $\hat{d}_i$ is computed analogously from $\{\hat{y}_{ij}\}$. We then report the Pearson correlation $\mathrm{Corr}(\{d_i\}, \{\hat{d}_i\})$, as in AART.

Baseline-specific Metrics AART uses exactly the three metrics above: annotator-level F1, global-level F1, and item-level disagreement correlation. We thus report all three for direct apples-to-apples comparison. AE instead evaluates each annotator's labels by computing (i) global accuracy over all (x_i, a_j) pairs and (ii) global-level F1 as defined above. Because AE does not release per-annotator predictions, we we are unable to use the annotator-level metrics for that baseline.

Trainable Parameters Lastly, we report the approximate number of trainable parameters for each model (our hypernetwork+adapters, AART, and AE). For AART and AE, we calculate these numbers based on their public source code, while for our method we report the number directly. Analyzing parameter counts helps us inquire into the parameter efficiency of each solution and assess how well each architecture can scale to the growing number of annotators to model and growing embedding space.

5 Results

In our experiments, the proposed hypernetwork-based architecture demonstrates generally strong performance across all datasets when compared to both AART and AE baselines, while using substantially fewer trainable parameters. Tables 2 and 3 report the mean and standard deviation over ten runs for each metric suite as defined in Section 4.3.

On the $\mathcal{D}_{MDA}$ dataset, our model achieves an annotator-level F1 of 70.24 ± 0.9 and a global-level F1 of 78.11 ± 0.2, thus surpassing both the AART baseline (annotator F1 69.72 ± 1.1, global F1 77.38 ± 0.4) and the AE baseline (global accuracy 75.14, global F1 73.60). This result indicates that our architecture is effective at capturing individual annotator tendencies and producing coherent overall predictions in this scenario.

For $\mathcal{D}_{RB}$, our approach again outperforms AART in annotator-level performance (73.81 ± 2.0 vs. 71.10 ± 3.2) and yields a competitive global-level F1 of 76.17 ± 1.8 (AART: 79.96 ± 1.9), suggesting that our hypernetwork is capable of modeling both base and rare annotator behaviors even when the dataset exhibits varied disagreement patterns.

Interestingly, on the more challenging $\mathcal{D}_{EPIC}$ dataset, our model gives lower annotator-level F1 (53.16 ± 1.6) and global-level F1 (65.11 ± 1.2) compared to AART's 59.67 ± 0.9 and 66.16 ± 1.4,

Dataset	*Single-task*	*AART*	*Ours*
		Annotator-level F1	
$\mathcal{D}_{MDA}$	66.80 ± 0.7	69.72 ± 1.1	$\mathbf{70.24 \pm 0.9}$
$\mathcal{D}_{EPIC}$	58.59 ± 1.9	$\mathbf{59.67 \pm 0.9}$	53.16 ± 1.6
$\mathcal{D}_{RB}$	68.61 ± 1.5	71.1 ± 3.2	$\mathbf{73.81 \pm 2.0}$
		Global-level F1	
$\mathcal{D}_{MDA}$	71.99 ± 0.6	77.38 ± 0.4	$\mathbf{78.11 \pm 0.2}$
$\mathcal{D}_{EPIC}$	60.23 ± 1.7	$\mathbf{66.16 \pm 1.4}$	65.11 ± 1.2
$\mathcal{D}_{RB}$	71.97 ± 1.7	$\mathbf{79.96 \pm 1.9}$	76.17 ± 1.8
		Item-level Disagreement Correlations	
$\mathcal{D}_{MDA}$	NA	0.37 ± 0.04	$\mathbf{0.42 \pm 0.02}$
$\mathcal{D}_{EPIC}$	NA	0.20 ± 0.06	$\mathbf{0.28 \pm 0.03}$
$\mathcal{D}_{RB}$	NA	0.54 ± 0.04	$\mathbf{0.66 \pm 0.03}$
		Trainable Parameters	
-	$124.3 * 10^6$	$\approx 124.9 * 10^6$	$\approx 5.6 * 10^6$

Table 2: This table shows the metrics of our model, averaged over 10 runs, against RoBERTa-based baselines from Mokhberian et al. 2024 (cols 1, 2; values as reported). Col. 2 features the configuration $\alpha > 0$; see the original paper for details. Along with average values, we report the standard deviation. The highest values are given in bold.

Dataset	*Single-task*	*AE*	*Ours*
		Global-level Accuracy	
$\mathcal{D}_{MDA}$	75.65	75.14	$\mathbf{80.11 \pm 0.2}$
$\mathcal{D}_{HS\text{-}Brexit}$	86.77	**87.03**	86.49 ± 0.8
		Global-level F1	
$\mathcal{D}_{MDA}$	73.26	73.60	$\mathbf{78.11 \pm 0.2}$
$\mathcal{D}_{HS\text{-}Brexit}$	**64.60**	60.36	58.30 ± 9.8
		Trainable Parameters	
-	$124.3 * 10^6$	$\approx 125.5 * 10^6$	$\approx 5.6 * 10^6$

Table 3: This table shows the metrics of our model, averaged over 10 runs, against RoBERTa-based baselines from Deng et al. 2023 (cols 1, 2; values as reported; std not reported). Col. 2 features the configuration E_a; see the original paper for details. The highest values are given in bold.

respectively. We believe this shortcoming stems from the higher complexity of data in EPIC,[3] which may require more specialized regularization strategies; still, our model attains a higher item-level disagreement correlation (0.28 ± 0.03 vs. 0.20 ± 0.06), thus showing that these two perspectivist metrics have a potential discrepancy.

On $\mathcal{D}_{\textit{HS-Brexit}}$, the comparison with AE yields a global accuracy of 86.49 (AE: 87.03) and a global F1 of 58.30 (AE: 60.36), which is only slightly below the baseline. Given the dataset size, it translates into a difference of $\tilde{8}$ misclassified instances.

Beyond model generalizability, `hypernetwork+adapters` solution demonstrates remarkably better parameter efficiency, as it requires only $\approx 5.6 * 10^6$ trainable parameters compared to $\approx 124.9 * 10^6$ for AART and $\approx 125.5 * 10^6$ for AE. This advantage is especially important within the strong perspectivist paradigm, where an ideal model should be able to scale up to hundreds of perspectives and, at the same time, take up a reasonable amount of memory and disk space.

6 Discussion

Two potentially advantageous aspects of the `hypernetwork+adapters` setting are (1) keeping the base model's weights intact and (2) adapting the model to all targets by training just one component. In what follows, we analyze why these aspects are especially relevant perspectivist learning.

Preserving the base model's original state could be beneficial for two reasons. Concerning inherently multi-purpose models, such as T5 (Raffel et al., 2020), it means that the base model's performance on other tasks will not be affected. This effect is especially important when dealing with large language models: they are often used for a wide range of tasks, and updating their weights can lead to catastrophic forgetting (Kirkpatrick et al., 2017) or degradation of their performance on previously learned tasks. For instance, if a model is fine-tuned for a particular task like hate speech detection, preserving the base model's original state ensures that its performance on other tasks remains intact.

Moreover, when adapting a model that has already been fine-tuned for some task irrespective of perspectives, our method can preserve this original, 'perspective-neutral' model. In a realistic setting of personalized content moderation, this model can be used as a fallback option. For example, if some user's view is not covered by the perspectivist model, the original, perspective-neutral model can be used to provide a fallback response.

Another benefit of modeling adapters and keeping the main classifier frozen is the mitigation of potential negative outcomes associated with perspectivist learning. In particular, when a RoBERTa model that is fully trainable is finetuned for perspectivist labels (as is the case in Deng et al. 2023 & Mokhberian et al. 2024), it is taught to associate the same text with varying labels (one to many). This can cause conflicting gradient steps leading the model away from the optimum. Incorporating annotator features through embeddings is intended to resolve this issue by converting the task back to a one-to-one association. However, it is unclear whether these features receive sufficient weight in the classifier to achieve that. In our framework, there is no discrepancy between inputs and outputs, as the only trainable component is the hypernetwork, which learns a one-to-one correspondence between annotators and their respective adapters. This approach does not expose the base model's weights to controversial targets and thus avoids the associated difficulties.

Dataset	*RoBERTa*	*LoRA*	*Ours*
	Trainable Parameters		
$\mathcal{D}_{\textit{MDA}}$	$334 * 125 * 10^6$	$334 * 6.6 * 10^5$	$5.6 * 10^6$
$\mathcal{D}_{\textit{EPIC}}$	$74 * 125 * 10^6$	$74 * 6.6 * 10^5$	$5.4 * 10^6$
$\mathcal{D}_{\textit{RB}}$	$819 * 125 * 10^6$	$819 * 6.6 * 10^5$	$5.9 * 10^6$
$\mathcal{D}_{\textit{HS-Brexit}}$	$6 * 125 * 10^6$	$6 * 6.6 * 10^5$	$5.3 * 10^6$

Table 4: Overview of trainable parameter counts required for fine-tuning a RoBERTa model to each perspective using all parameters, low-rank adaptation, and our hypernetwork respectively ($r = 2, \alpha = 32$).

Using a hypernetwork to learn adapter weights offers an additional advantage. Training a separate LoRA-style adapter for each annotator is possible but involves training a large number of parameters that grows substantially with each added perspective. As shown in Table 4, the costs of adapting RoBERTa-base to every perspective in our datasets are substantial. In extreme cases, such as $\mathcal{D}_{RB}$ with 819 annotators, the required parameters exceed not only the hypernetwork's but also RoBERTa's total parameter count. This results in impractical

[3] For example, as per Casola et al. (2024), gpt-3.5-turbo yields an F1 of 48.1 on EPIC in zero-shot settings; this differs from its zero-shot performance on other language understanding tasks, such as sentiment analysis or natural language inference (F1s of 91.13 and 67.87 respectively, Ye et al. (2023)).

memory and disk space requirements, rendering the separate adapter approach infeasible.

In contrast, our method stands out as the more cost-effective solution in all settings, except for $\mathcal{D}_{HS\text{-}Brexit}$, which has an exceptionally small number of annotators[4]. Like other approaches that make use of annotator embeddings, adding a new annotator requires $1 \times hdim$ parameters + rescaling the model to a new embedding. Given the small size of the hypernetwork, all of this sums up to just $2 \times hdim$, allowing for very cheap scaling to additional perspectives. This property makes it especially attractive for perspectivist learning.

Conclusion

In this work, we have investigated a `hypernetwork+adapters` architecture for perspectivist learning on subjective classification tasks. Our experiments show that, while this model does not uniformly exceed the latest perspective-aware baselines, it achieves superior performance on several perspectivist datasets, most notably $\mathcal{D}_{MDA}$ and $\mathcal{D}_{RB}$. It also obtains higher item-level disagreement correlations even when mean F1 is lower, as on $\mathcal{D}_{EPIC}$. Notably, our approach requires only about 5.6×10^6 trainable parameters, about 4.5% with respect to over 124×10^6 in the competing methods; thus, it offers a considerable advantage in parameter efficiency.

Taken together, these findings suggest that the `hypernetwork+adapters` design is a promising solution within the strong perspectivist paradigm, even if further work is needed to let it scale equally well to all available tasks and datasets.

Limitations

One important limitation of the proposed hypernetwork architecture is that it takes more time for inference, as it follows a two-stage procedure where it first separately predicts the LoRA weights of each adapted layer and then applies these during classification. This flow leads to both training and evaluation taking longer than in the case of regular classifier models ($\approx$ 0.25 iterations/sec. vs. 4 iterations/sec.). We argue, however, that this shortcoming does not outweigh the advantage in parameter efficiency. The decreased memory consumption allows for more parallel training jobs to be scheduled simultaneously, thus compensating for lower throughput.

A further possible limitation is that we only use annotator IDs as annotator features. This strategy does not permit our system to scale to new perspectives if we need to make predictions for an unseen annotator. We see two ways to address this limitation. First, in this paper, we do not inquire into how well hypernetworks work with sociodemographic features. However, they can still be trivially integrated, possibly mitigating this issue. A further way to tackle this problem could consist in finding an annotator in the existing annotator pool that is most similar to the unseen one and assuming their perspective. Similarity of annotators could be approximated from the said sociodemographic features.

Finally, we acknowledge that our evaluation of the proposed architecture is not exhaustive, and we could overlook some shortcomings of our model. However, we find it sufficient to judge how well it compares to the competitor models. Additionally, we hope to scrutinize it more when more perspectivist models and datasets are released.

[4] In order to verify that separately-trained LoRA adapters do not significantly outperform our model, we conduct an experiment on $\mathcal{D}_{HS\text{-}Brexit}$. We report the outcomes in Table 5; they suggest that training separate adapters does not surpass our approach.

Acknowledgments

We thank the anonymous reviewers and the NLP group at the University of Utrecht for their helpful feedback on this work. Generative tools, such as LLAMA-3 (Grattafiori et al., 2024), were used to proof-read this text. This study is funded by NWO through an AINed Fellowship Grant NGF.1607.22.002 and supported by project 'Dealing with Meaning Variation in NLP'.

References

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

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Connor Baumler, Anna Sotnikova, and Hal Daumé. 2023. Which examples should be multiply annotated? active learning when annotators may disagree.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings*

of the AAAI Conference on Artificial Intelligence, 37(6):6860–6868.

Silvia Casola, Simona Frenda, Soda Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and 1 others. 2024. Multipico: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. 2025. Text-to-lora: Instant transformer adaption. *arXiv preprint arXiv:2506.06105*.

Christopher Clarke, Yuzhao Heng, Lingjia Tang, and Jason Mars. 2024. Peft-u: Parameter-efficient fine-tuning for user personalization. *arXiv preprint arXiv:2407.18078*.

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.

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. *Preprint*, arXiv:2305.14663.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

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.

Preni Golazizian, Ali Omrani, Alireza S. Ziabari, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *ArXiv*, abs/2402.14101:null.

Faustino Gomez and Jürgen Schmidhuber. 2005. Evolving modular fast-weight networks for control. In *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, pages 383–389, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Ha, Andrew Dai, and Quoc V. Le. 2017. Hypernetworks. *ICLR*, 1(8):103–121.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, and 1 others. 2022. Hyperprompt: Prompt-based task-conditioning of transformers. In *International conference on machine learning*, pages 8678–8690. PMLR.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA. USENIX Association.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Max Müller-Eberstein, Dianna Yee, Karren Yang, Gautam Varma Mantena, and Colin Lea. 2024. Hypernetworks for personalizing ASR to atypical speech. *Transactions of the Association for Computational Linguistics*, 12:1182–1196.

Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. 2023. Hypertuning: Toward adapting large language models without back-propagation. In *International Conference on Machine Learning*, pages 27854–27875. PMLR.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2023. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6527–6536.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

O. O. Sarumi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus considerations for annotator modeling and scaling. *ArXiv*, abs/2404.02340:null.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. Value profiles for encoding human variation. *Preprint*, arXiv:2503.15484.

Derya Soydaner. 2020. Hyper autoencoders. *Neural Processing Letters*, 52(2):1395–1413.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491.

Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, and Da-Cheng Juan. 2021. Hypergrid transformers: Towards a single model for multiple tasks. In *International conference on learning representations*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-X: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 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, Online. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, and 1 others. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

A Separate LoRA training.

Dataset	*Baseline*	*LoRA*	*Ours*
Global-level Accuracy			
$\mathcal{D}_{\textit{HS-Brexit}}$	86.77	77.50	86.49
Global-level F1			
$\mathcal{D}_{\textit{HS-Brexit}}$	64.60	53.02	58.30

Table 5: This table reports the metrics of our model against the performance of 6 separate LoRA adapters (one per each annotator perspective) on HS-Brexit. The training parameters for the adapters copied those of the main experiment, the exception being an increased learning rate ($5\mathrm{e}{-5}$) and an increased epoch count (10). We report the mean results per 10 runs.