Voices in a Crowd: Searching for clusters of unique perspectives

Anonymous ACL submission

Abstract

Fine-tuned models have been shown to reproduce underlying biases existing in their training data, which is by default the majority perspective. While this process has been shown to minimise minority perspectives, proposed solutions either fail to preserve nuances present in the original data, or are based on strong a-priori assumptions about annotators that when used can bias model training. We propose an approach that trains models purely in an annotator demographic-agnostic manner, extracts latent embeddings informed by annotator behaviour during training, and clusters annotators based on their behaviour over the respective corpus. Resulting clusters are subsequently validated post-hoc via internal and external validative quantitative metrics, as well as our resulting qualitative analysis. Our results explain the strong generalisation capability of our framework, indicated by resulting clusters being adequately robust, while also capturing minority perspectives based on different demographic factors throughout two distinct datasets.[1]

Content Warning: This document contains and discusses examples of potentially offensive and toxic language.

1 Introduction

Supervised training of Machine Learning (ML) and Natural Language Processing (NLP) models is rooted in the presupposition that for every example in a dataset, a ground truth, also known as a gold label, exists. This allows for an objective measure of success; a model has learned the underlying patterns from the data if its prediction for an example is congruent with the ground truth (Hettiachchi et al., 2021).

However, the concept of a single ground truth per item can be particularly challenging to assess in subjective tasks in cases of pervasive annotator disagreement persisting throughout a dataset (Uma et al., 2022, 2021). While such disagreement can be indicative of task difficulty or semantic ambiguity

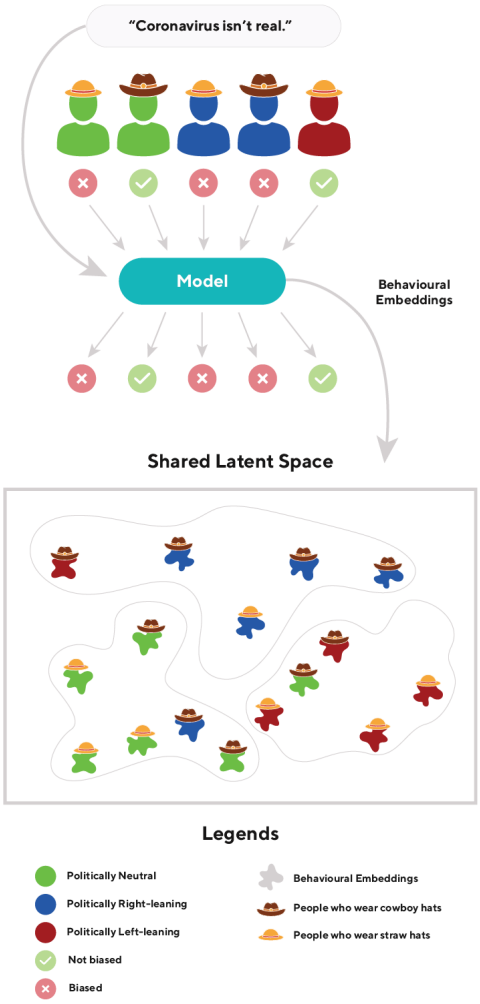


Figure 1: Models are trained through text examples and annotations, with models learning to predict the unique perspectives of each annotators without any further annotator metadata. Decoder hidden states are subsequently used to cluster annotator opinions on a given corpus through unsupervised methods to find emergent groups of unique minority perspectives not fully captured via sociodemographic information.

[1]We will release the codebase on GitHub upon acceptance.

(Wang et al., 2021; Jiang and Marneffe, 2022; Sandri et al., 2023), it can also indicate stable and conflicting intra-annotator perspectives (Basile et al., 2020; Abercrombie et al., 2023).

Efforts on how to best deal with clashing perspectives include changes to both how models are trained and evaluated. Regarding model training, recent approaches have proposed introducing demographic-level labels alongside annotations to improve a model's representative capacities and generalisability towards a target minority group present in the dataset (Fleisig et al., 2023; Gupta et al., 2023; Beck et al., 2023). For evaluation, the consensus seems to be rejecting gold labels in favour of more representative metrics or methodologies, e.g., distributional labels that capture per-item disagreements which allow for degrees of confidence on predictions (Leonardelli et al., 2023).

Although preferable to gold labels, such solutions are still vulnerable to collapsing multiple viewpoints into a minority–majority distribution per item (Gordon et al., 2022). While recent approaches have explained the importance of capturing and preserving distinct perspectives as-is (Vitsakis et al., 2023; Cabitza et al., 2023), it remains unclear as to which modelling and training choices should be employed to best do so.

Our Contributions We introduce a framework that evaluates how training choices affect a model's ability to preserve distinct perspectives in a dataset. We employed six distinct modelling architectures on two heterogeneous datasets, all within the context of annotation of political bias. Models were trained to predict individual annotations per annotator in a demographic-agnostic manner, i.e., no annotator information was provided as input to the models. Extracted the latent embeddings of each model were used them to *cluster annotations* into groups through unsupervised learning.

We show that such models can correctly identify perspectives learned through similarities between annotator behaviours (their annotation patterns) as captured in latent spaces without the need for further information. Crucially, we validated the created clusters by matching the demographic information of annotators post-hoc, and conducted an in-depth qualitative analysis of the clusters themselves. Since the models are trained without any demographic information, our method uniquely allows us to explain the impact of demographics on different datasets without constraints, evidenced by the creation of clusters based solely on different demographics that emerged organically throughout distinct datasets.

2 Related Work

Dealing with disagreements Aggregating annotator disagreements into a single gold label per item can improve model performance (Nguyen et al., 2017). However, such approaches also imprint the resulting model with a simplified and reduced view of the minority perspectives present in the data (Gordon et al., 2022), leading to further erasure of underrepresented minorities of annotators (Prabhakaran et al., 2021).

One solution is to supplement gold labels with silver labels, i.e., distributional per-item labels that measure disagreement amongst annotators (Leonardelli et al., 2023; Uma et al., 2022, 2021). While this approach allows for the identification of controversial items in datasets (Fornaciari et al., 2022), it fails to capture stable inter-annotator disagreements throughout the dataset that could provide insight as to why disagreement occurs beyond an item-by-item scale (Abercrombie et al., 2023).

Demographics and annotator Bias Bias introduced through annotations is an established phenomenon (Hovy and Prabhumoye, 2021; Garrido-Muñoz et al., 2021; Blodgett et al., 2020; Geva et al., 2019). Individual annotator characteristics such as age (Al Kuwatly et al., 2020), gender (Stanczak and Augenstein, 2021; Biester et al., 2022), or political orientation (Baly et al., 2020; Sap et al., 2021), have all been shown to impact annotator behaviour, and consequently, model performance in classification tasks.

Proposed solutions have attempted to incorporate information about annotator beliefs (Röttger et al., 2021; Davani et al., 2023), or demographics (Fleisig et al., 2023; Gupta et al., 2023) into the training pipeline to allow learning of patterns between annotations and in-group tendencies. While incorporation of such information can seemingly improve model performance in specific tasks (Welch et al., 2020), evidence suggests that such results might not be generalisable across datasets (Lee et al., 2023). Since demographics are not necessarily predictive of underlying annotator beliefs (Hwang et al., 2023; Beck et al., 2023), there is a strong need for models that capture annotator perspectives without the need for a priori assumptions.

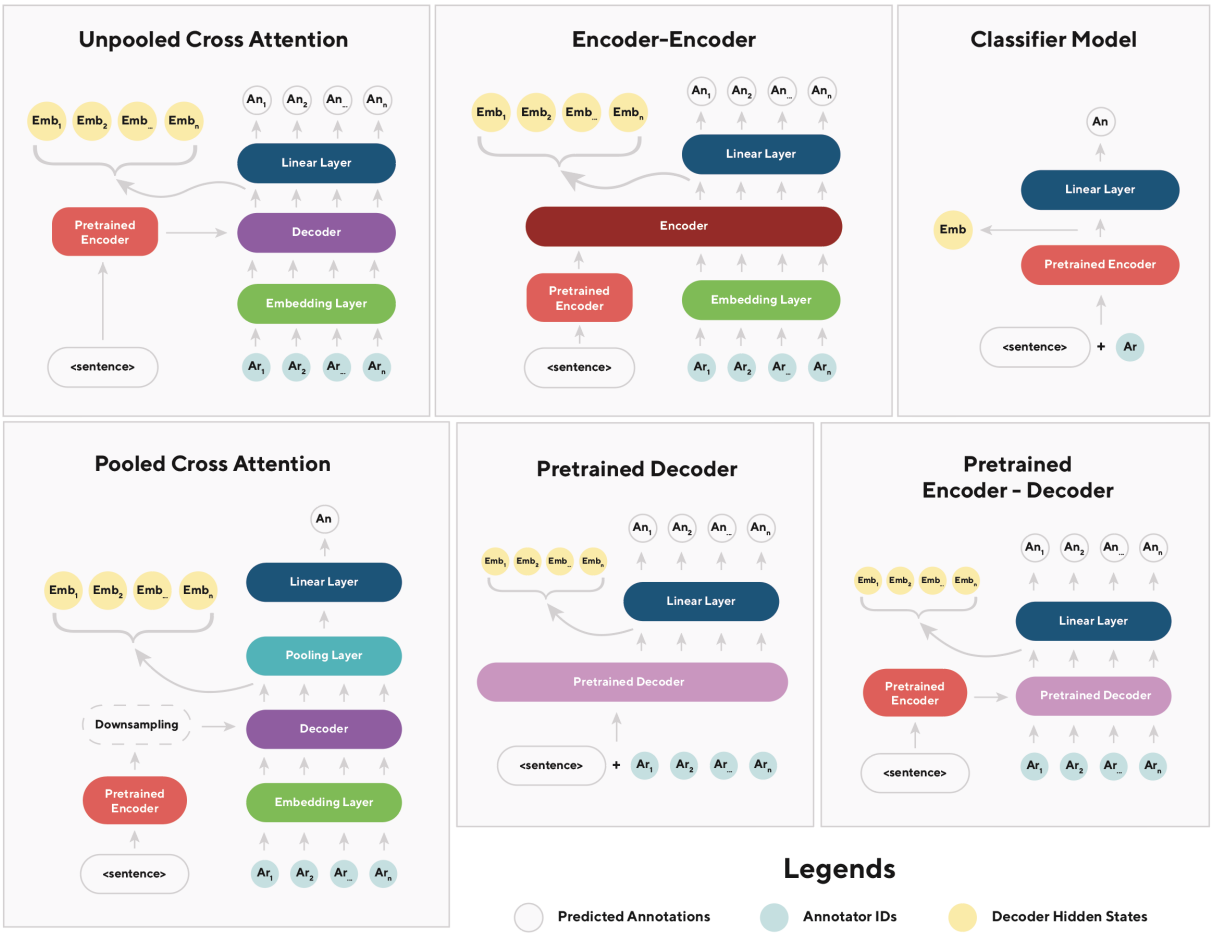


Figure 2: Training component: 6 modelling architectures for extracting decoder hidden states (denoted with a yellow circle as Emb_n) used as input for the Clustering component.

Unsupervised learning and clustering of attitudes Unsupervised learning has been used to identify emergent themes within corpora via clustering of latent textual embeddings (Sevillano et al., 2007; Meng et al., 2022). Dhillon and Modha (2001) explain that using textual elements (i.e., word embeddings) as features in a high-dimensional latent space allows for clustering based on inter-dimensional similarities. However, fine-tuning pre-trained language model embeddings tend to produce embeddings that are anisotropic and anisometric (Rajaee and Pilehvar, 2021; Xu and Koehn, 2021) in nature; when paired with their high dimensionality, clustering via distance-based metrics becomes challenging.

Nevertheless, recent findings indicate that although isotropy still exists, distance metrics can still be employed after employing dimensionality reduction methods (Mu et al., 2017). More specifically, since similar contextual embeddings follow a spiral-band, or Swiss-roll manifold shape (Cai et al., 2020), we can use appropriate dimensionality reductions to then analyse relationships between features through Euclidean distance-based metrics (McInnes et al., 2018).

3 Experimental Methodology

Our approach consists of two components. First, we explore several modelling choices (Section 3.2) for supervised fine-tuning to predict each annotator's individual annotation for a given example. This ensures we preserve the unique perspective without biasing the model by providing additional information (Vitsakis et al., 2023). Each model is fine-tuned separately for each of our chosen datasets. Secondly, we cluster the resulting latent spaces from each model which have been informed by each annotator's opinion of the text during training. These embeddings are then processed through one of two dimensionality reduction techniques (Section 3.3) before being clustered using K-Means (MacQueen et al., 1967; Pedregosa et al., 2011).

3.1 Datasets

All datasets used in our experiments contain annotator demographics such as personal political leaning, age, and education level.

Media Bias Annotation Dataset (MBIC) (Spinde et al., 2021a,b) comprises sentences from media articles that may contain political bias from news outlets across the political spectrum (e.g., Fox News, MSNBC, etc.) covering 14 potentially divisive topics (e.g., gender issues, coronavirus, the 2020 American election). 784 crowd-sourced annotators labelled sentences on whether they consider them to contain bias. Demographics of the dataset were slightly skewed throughout dimensions such as political ideology (44.3% left learning, 26.7% right-leaning, 29.1% center).

Global Warming Stance Dataset (GWSD) (Luo et al., 2020) contains opinions of varying intensities on the subject of global warming, gathered from news outlets of varied political leanings (e.g., The New York Times, and Breitbart). 398 annotators labelled each sentence with whether they agreed, disagreed, or were neutral. Demographic skew of this dataset mirrored that of MBIC in self reported political affiliation (46% Democrat, 21.2% Republican, 28.8% Independent, 4% Other).

3.2 Training component

We evaluated the performance of six distinct modelling architectures, each trained through a different combination of our inputs as seen in Fig. 2. For a given text sample in a dataset, $\mathbf{x} \in \mathbf{X}$, each model predicts the individual annotation of each annotator $p_\theta(\mathbf{y}|\mathbf{x})$ where $\mathbf{y} = (y_1, \ldots, y_K)$, and K is the total number of unique annotators within the dataset.

Unpooled Cross Attention This model uses a pretrained T5 encoder-decoder model (Raffel et al., 2020) where the encoded text and the embedded encoded annotator unique identifiers are fed through a decoder layer which predicts each annotator's annotation as a sequence.

Encoder-Encoder Inspired by multi-modal approaches which leverage distinct modalities through either text and vision (Tan and Bansal, 2019; Singh et al., 2022; Agarwal et al., 2020), this architecture treats text and annotators as separate modalities. The encoded text and embedded annotator unique identifiers are fed through a cross-modality encoder (Raffel et al., 2020) to predict the annotation of each annotator.

Classifier Model This architecture uses a transformer-based encoder-only classifier as a backbone model, i.e., BERT (Devlin et al., 2018) for GWSD, and RoBERTa (Liu et al., 2019) for MBIC. We simply concatenate the text with a unique annotator identifier and predict each label independently.

Pooled Cross Attention This model is based on Sullivan et al. (2023)'s approach, which showed strong results during the 2023 Learning With Disagreements (LeWiDi) shared task (Leonardelli et al., 2023) in predicting annotator disagreement. Largely similar in structure to 'Unpooled Cross Attention', it also uses a T5 encoder-decoder model as the backbone. Then the encoded text embedding dimension gets reduced through downsampling as previous research has indicated possible benefits in salience of encoded features (Schick and Schütze, 2019; Dhingra et al., 2018; Holzenberger et al., 2018). Finally, decoder outputs are pooled (Reimers and Gurevych, 2019) into a shared latent space that is used to predict an aggregated annotation for each batch.

Pretrained Decoder This architecture uses a pre-trained GPT-2 decoder (Radford et al., 2019) that receives as input the concatenated text and annotator identifiers of the form "`<text> [SEP] <Ann1> [SEP] ... <AnnN>`" and predicted the annotation for each annotator.

Pretrained Encoder-Decoder This architecture is similar to 'Unpooled Cross Attention'. The model uses a pre-trained T5 encoder-decoder instead (Raffel et al., 2020); the only difference is that the unique annotator identifiers were embedded through the decoder tokenizer of the T5 model itself, to be able again to predict each annotator's annotation, autoregressively.

Metrics Since both datasets have a fairly unbalanced distribution of labels we report precision, recall, and F1 score. Average pairwise cosine similarity between decoder hidden states of predicted annotations were also procured. Since this metric shows how dense the decoder latent state is by the end of training; a lower score generally correlates with better clustering performance.

Results Table 1 summarises the results for the Training component experiments. For the GWSD

	F1 Score ↑	Precision ↑	Recall ↑	Avg. Pairwise Similarity ↓
GWSD Dataset				
Cross Attention	**0.65**	0.64	**0.65**	0.14 ± 0.07
Pooled Cross Attention	0.19	0.14	0.33	0.54 ± 0.13
Encoder-Encoder	0.63	0.66	0.62	0.15 ± 0.11
Classifier Model	0.63	**0.67**	0.61	0.81 ± 0.14
Pretrained Decoder	0.62	0.64	0.61	0.66 ± 0.08
Pretrained Encoder-Decoder	0.19	0.28	0.34	0.95 ± 0.02
MBIC Dataset				
Cross Attention	**0.72**	**0.72**	**0.72**	0.22 ± 0.05
Pooled Cross Attention	0.43	0.47	0.41	0.70 ± 0.06
Encoder-Encoder	**0.72**	**0.72**	**0.72**	0.21 ± 0.06
Classifier	0.38	0.3	0.5	1.00
Pretrained Decoder	0.63	0.65	0.63	0.75 ± 0.07
Pretrained Encoder-Decoder	0.71	0.71	0.71	0.74 ± 0.25

Table 1: Overall performance (Precision/Recall, and F1 score) for the training component of our framework (6 modelling architectures) on MBIC and GWSD for the task of individual annotator prediction. We also report the average pairwise cosine similarity across decoder hidden states for every model; lower score indicates greater variety in representation which correlates with better clustering performance.

dataset, the Cross Attention architecture performed best overall, while for the case of the MBIC dataset, the Cross Attention and Encoder-Encoder architectures resulted in the highest F1 Score. The Classifier Model was the worst performing model for the MBIC dataset and had an average pairwise similarity of 1, indicating that the decoder hidden states are near-identical. Similarly, one of the worst performing models for the GWSD dataset also has high pairwise similarity across the decoder hidden states. Models with the highest F1 score across both datasets also have a low average pairwise similarity across the decoder hidden states, which indicates that the latent state of the models are less dense.

3.3 Clustering component

Next, we move on to clustering the decoder hidden states of the annotation embeddings. For the remainder of the paper, we used the outputs of the 'Encoder-Encoder' model as it has the highest F1 scores and on average lowest pairwise similarities across both datasets (see Table 1). Following the discussion in Section 2, we perform dimensionality reduction first before proceeding to obtain the clusters.

Dimensionality Reduction We experimented with the following dimensionality reduction techniques: a baseline of no dimensionality reduction, Principal Component Analysis (PCA; a linear combination of components), and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP; a non-linear transformation algorithm) (McInnes et al., 2018). Both PCA (Sia et al., 2020; Gupta et al., 2019), and UMAP (Cai et al., 2020; Ait-Saada and Nadif, 2023; George and Sumathy, 2023) have been previously shown to improve feature representation in high-dimensional latent spaces leading to improved clustering performances.

Clustering Techniques We used K-means (MacQueen et al., 1967; Pedregosa et al., 2011) to cluster the behavioural embeddings resulting from the different options of dimensionality reduction. K-means exhibits robustness and performs well when clustering features from high-dimensional latent spaces created from text (Song and Park, 2007; Rashid et al., 2020; Ahmed et al., 2022), especially when paired with PCA (Hosseini and Varzaneh, 2022), or UMAP (Allaoui et al., 2020).

Metrics We used two **internal validation** metrics namely, *Silhouette* (Rousseeuw, 1987; Pedregosa et al., 2011) and *Davies-Bouldin Index* (Davies and Bouldin, 1979; Pedregosa et al., 2011); both are used to assess average similarity scores between clusters. Silhouette assesses intra cluster separatation and is bound between -1 and 1, with 1 being the best possible score, with the threshold for moderate clusters being being 0.5 (Shahapure and Nicholas, 2020; Lengyel and Botta-Dukát, 2019).

	# Clusters	Davies-Bouldin Index↓	Silhouette ↑	Purity-Political ↑	Purity-Education ↑
MBIC					
Cross Attention					
No dim. reduction	19	6.35	0.02	**0.71**	**0.71**
w/ PCA	14	1.10	0.25	0.36	0.43
w/ UMAP	19	**0.50**	**0.53**	0.50	0.41
Pooled Cross Attention					
No dim. reduction	19	3.03	0.06	0.42	0.48
w/ PCA	19	**0.51**	**0.53**	0.44	0.51
w/ UMAP	10	**0.51**	**0.53**	**0.49**	**0.55**
Encoder-Encoder					
No dim. reduction	19	6.93	0.01	0.41	0.46
w/ PCA	19	**0.49**	**0.54**	**0.53**	0.43
w/ UMAP	19	**0.49**	0.53	0.51	**0.48**
Classifier Model					
No dim. reduction	5	6.37	0.04	**0.45**	**0.50**
w/ PCA	13	**0.50**	**0.55**	0.43	0.44
w/ UMAP	18	0.52	0.50	0.38	**0.50**
Pretrained Decoder					
No dim. reduction	19	2.86	0.06	0.47	0.47
w/ PCA	19	0.50	0.53	0.44	0.52
w/ UMAP	19	**0.49**	**0.55**	**0.50**	**0.53**
Pretrained Encoder-Decoder					
No dim. reduction	5	1.70	0.16	0.47	0.48
w/ PCA	19	**0.48**	**0.55**	0.46	0.46
w/ UMAP	14	0.49	0.56	**0.53**	**0.49**

Table 2: Overall performance through internal (Davies-Bouldin Index, Silhouette) and external (Purity Political / Education) validity measures for the clustering component of our framework. Intra-cluster separation indicative of better overall clustering performance indicated by higher Silhouette and lower Davies-Bouldin scores. External validity, measured via inter-cluster purity, indicated by higher purity scores

The Davies-Bouldin Index is also a measure of intra-cluster dissimilarity, as indicated by the lowest possible score with a lower bound of 0 (Idrus, 2022; Kärkkäinen and Fränti, 2000).

We used *Purity* to assess the **external validity** of clusters. Purity measures the internal consistency of assigned labels within a cluster. It has been previously used to evaluate whether a cluster is prototypical (i.e., representative) across provided labels within a dataset (Christodoulopoulos et al., 2010). In our case, it allows us to automatically assess whether a cluster emerging from annotator behaviours during training is linked to any of the annotator labels (e.g., a cluster with high right-leaning political consistency) and thus is indicative of a distinct perspective.

The labels used were based on political orientation and education levels. Each dataset collected the labels slightly differently: GWSD represents political party affiliation as a categorical variable (i.e., "democrat", "other", "republican" etc.), while MBIC uses a range of values between -10 and 10.

3.4 Quantitative Cluster Validation Results

Optimal cluster numbers were automatically calculated through a series of hyperparameter sweeps to minimise the Silhouette score (see Appendix A for more information). Table 2 shows the best performing clusterings for MBIC; results for GWSD can be found in Table 12 (Appendix B.2).

Internal Validity Metrics Overall, the choice of dimensionality reduction significantly impacted the quality of the resulting clusters as indicated through our internal validation metrics. On the GWSD dataset, UMAP and PCA marginally outperformed each other in terms of Silhouette and Davies-Bouldin Index scores respectively, while in the MBIC dataset, UMAP outperformed PCA on both internal validation metrics. No dimensionality reduction consistently resulted in poor cluster

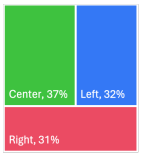
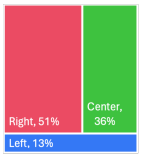
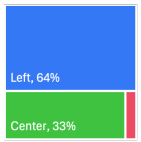
Dataset/Cluster No.	Examples	Bias Label	Distribution
MBIC -1	British Olympic swimmer Sharron Davies also slammed the concept of transgender athletes.	✓	
	BBC Presenter Gabby Logan has said that it is not fair that transgender women can compete in sport alongside biologically female women.	✓	
	BBC Presenter Gabby Logan has said that it is not fair that transgender women can compete in sport alongside biologically female women.	✗	
MBIC -7	Trump — who has been criticized for painting an overly rosy picture of the outbreak, often contradicting his own health officials - insisted on Friday that his administration was "magnificently organized" and "totally prepared" to address the virus.	✓	
	Google declined to offer details beyond Huntley's tweets, but the unusually public attribution is a sign of how sensitive Americans have become to digital espionage efforts aimed at political campaigns.	✗	
	At least 25 transgender or gender-nonconforming people were killed in violent attacks in the United States last year, according to the Human Rights Campaign, which has been tracking anti-trans violence since at least 2015.	✓	
MBIC -8	Though conservatives try to demonize Ocasio-Cortez an Omar, their actual policy views are perfectly mainstream. The New York lawmaker proposed a 70 percent tax on top incomes — a view backed by public opinion and many well-respected economists.	✗	
	British Olympic swimmer Sharron Davies also slammed the concept of transgender athletes.	✗	
	At least 25 transgender or gender-nonconforming people were killed in violent attacks in the United States last year, according to the Human Rights Campaign, which has been tracking anti-trans violence since at least 2015.	✓	

Table 3: Analysis of clusters on the MBIC dataset with the Encoder-Encoder architecture and UMAP dimensionality reduction. We report the cluster number, representative examples of the cluster, and their paired annotation (✓ for perceived bias, ✗for no perceived bias). We also show the distribution of annotator characteristics which is indicative of the prototypical nature of each cluster.

quality as indicated by low Silhouette (Shahapure and Nicholas, 2020; Idrus, 2022) scores. Finally, PCA resulted in a smaller average number of optimal clusters in the MBIC dataset compared to UMAP while the opposite is true for the MBIC dataset. We report the averages of internal validity scores across different dimensionality techniques in Appendix B.1.

External Validity Metrics Individual purity scores assigned to each cluster can help us interpret whether clustering of behavioural embeddings resulted in clusters prototypical of the demographics of our annotators. Purity scores of all created clusters per method were averaged, with the resulting scores shown in Table 12. However, since we are interested in finding prototypical clusters potentially small in size, we manually inspected several of them to assess the efficacy of our framework.

3.5 Qualitative Cluster Validation

3.5.1 MBIC Dataset

Table 3 shows the results of a case study of individual clusters resulting from K-means using UMAP on the behavioural embeddings from the Encoder-Encoder model on the MBIC dataset. We found emerging clusters with political orientation being their most salient feature. Any cluster that significantly deviated from the dataset's original label distribution (described in Section 3.1) was considered a potentially prototypical cluster. We pick three clusters (out of a single K-means run) and discuss them below:

Cluster 1 This cluster is a prototypical example of a consensus amongst annotators. Following similar trends to the original label distribution of the data (44.3%, 29.1%, 26.7% for left, center, and right political lean), the cluster's distribution is more even. Such clusters often contain different labels for the same sentences, while there is also no strong emerging effect from collected labels.

Cluster 7 This is a minority cluster, with distribution of labels indicating that this cluster is primarily formed through opinions that are right-leaning. While Item 1 is expectantly labelled as 'bias', Item 3 contains no obvious biased words, although it comes from an obvious place of concern for a marginalised minority.

Cluster 8 This is a prototypical example of a majority dominant cluster. Such clusters are populated by the opinion of the original dataset's distributional majority label although with a much heavier skew, indicating a stable and consistent behaviour of the group. Labelling distribution of this cluster is expected to be populated by left-leaning views and indeed sentences that were previously labelled as

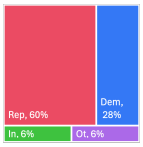
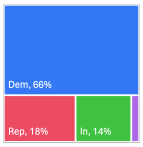
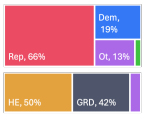
Dataset/Cluster No.	Examples	Agreement Label	Distribution
GWSD - 9	The early 21st-century drought that afflicted Central Asia is the worst in Mongolia in more than 1,000 years, and made harsher by the higher temperatures consistent with man-made global warming.	✓	
	Climate change means the end of shopping.	∼	
	The oil sands are responsible for just 0.001 percent of global greenhouse emissions	∼	
GWSD - 2	There is a connection between human activity and an assumptive change in global climate.	✓	
	Hiring a White House "climate change czar" would be a good idea.	✓	
	Scaring young people young people into believing that climate change is going to kill young people is child abuse.	✗	
GWSD - 5	The oil sands are responsible for just 0.001 percent of global greenhouse emissions	✓	
	This could mean that current I.P.C.C. model predictions for the next century are wrong, and there will be no cooling in the North Atlantic to partially offset the effects of global climate change over North America and Europe.	✓	
	Eco-towns could provide an inspiring blueprint for low-carbon living	✗	

Table 4: Analysis of clusters on the GWSD dataset using same parameters as the MBIC dataset, and results are shown in a similar fashion (✓agree with the statement, ✗for disagree and ∼ for neutral). Distribution of annotator characteristics is provided.

biased in non-left-leaning clusters (Item 1 of Cluster 1, and Item 3 of Cluster 7), were consistently found to not be labelled as such.

3.5.2 GWSD Dataset

Table 4 shows a case study on the GWSD dataset, under the exact same experimental conditions.

Cluster 9 This is a prototypical example of a minority cluster, as indicated by the differences in the distribution of the minority label between the cluster (60% and the original data (21%). Opinions expressed in the cluster were generally agreeable about climate-changing effects, we didn't find any more agreement with more politically charged statements.

Cluster 2 This is a majority-dominant cluster. Opinions that could be perceived as more political were found to be more common (Item 2), with some strong examples (Item 3).

Cluster 5 An example of a minority within a minority perspective. Opinions are over-represented by two minority labels, the "republican" in terms of political affiliation, and that of the "higher degree" in terms of education level (8.4% label representation in the original dataset). Opinions showed fewer "neutral" responses and were generally indicative of a well-informed audience, explicitly agreeing with more technical items such as Item 2 and especially Item 1, which received mostly "neutral" scores in other clusters (e.g., Cluster 9).

3.6 Summarisation of qualitative results

The results of Sections 3.4 and 3.5 showcase the generalisability capacities of our framework: our models produce embeddings that can be clustered based on behavioural patterns that capture perspectives indicative of population sample minorities. The three types of clusters found in GWSD (minority, majority, and minority within a minority) paired with the inspections of the sentence-annotation pairs validate our claims.

4 Conclusion

In this work, we propose a novel framework that can be used to find underlying minority perspectives in data. Six distinct model architectures were trained on a classification task while not being provided with any annotator meta-data to avoid biasing their training. Subsequently, the decoder hidden layers of trained models were passed through various methods of dimensionality reductions (UMAP and PCA), with the resulting embeddings used to create clusters through an unsupervised algorithm (K-means).

The resulting clusters were adequately separated according to internal validative metrics. Manual inspection of clusters produced by our best-performing models showcased the ability of our framework to capture perspectives as shown by three distinct types: clusters representative of a minority, a majority, and clusters that captured multiple minority labels, a minority within a minority.

Limitations and Ethical Considerations

4.1 Internal validity and dimensionality reduction

A current limitation of the model is the ability to automatically assess which methodologies perform without manual inspection. As shown in Tables 2 and 11 while internal validation scores *can* be indicative of well defined clusters of minority perspectives, they are not necessarily so. While PCA was only marginally outperformed by UMAP in terms of internal validation scores, the distributions of labels in the clusters resulting from a PCA dimensionality reduction were minimally different when compared to label distributions present in the original data.

Manual inspection of the clusters indicated that clusters were formed around the most salient features discovered during training, namely the unique annotator tokens, or the inter-sentence similarities. A possible reason for this phenomenon could be that PCA reduces dimensionalities to the most salient principal components, which have been shown to not be good conducive to clustering based on contextual features in large language models (Cai et al., 2020). Interestingly, this phenomenon can also be reproduced with UMAP when instructing the model to focus on finding clusters based on local and not overarching features (McInnes et al., 2018). A possible solution to this issue is offered by (Mu et al., 2017), which explain that removal of the top principal components results in more salient representations, and thus could improve clustering performance.

4.2 Labels and further marginalisation of minorities

Our model uses labels procured during data gathering to validate emergent clusters. However, the labelling gathering process can potentially be an erasing process towards minorities in and of itself (Hovy and Prabhumoye, 2021; Chandrabose et al., 2021). In our case we encountered this limitation with the GWSD dataset (Luo et al., 2020), which collected categorical labels about political affiliation of participants. Beyond the three primary labels ("Democrat", "Independent", "Republican"), the rest were aggregated into the "other" label. This resulted in a minority so small that our clustering methodology could not adequately disentangle.

The labelling process can also further discriminate against socially marginalised minorities by not providing options consistent with an individual's identity (Chandrabose et al., 2021; Jo and Gebru, 2020). A possible solution would be to then validate the content of our clusters through a method not based on collected data such as sentiment analysis, which has been previously used to classify politically opinions on charged data (Dorle and Pise, 2018; Kazienko et al., 2023; Ansari et al., 2020).

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? *arXiv preprint arXiv:2005.07493*.

Majid Hameed Ahmed, Sabrina Tiun, Nazlia Omar, and Nor Samsiah Sani. 2022. Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*, 13(1):342.

Mira Ait-Saada and Mohamed Nadif. 2023. Is anisotropy truly harmful? a case study on text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203.

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

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In *International conference on image and signal processing*, pages 317–325. Springer.

Mohd Zeeshan Ansari, Mohd-Bilal Aziz, MO Siddiqui, H Mehra, and KP Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia Computer Science*, 167:1821–1828.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. *arXiv e-prints*, page arXiv:2309.07034.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 10–19.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2020. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Aravindan Chandrabose, Bharathi Raja Chakravarthi, et al. 2021. An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Inderjit S Dhillon and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42:143–175.

Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. 2018. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*.

Saurabh Dorle and Nitin Pise. 2018. Political sentiment analysis through social media. In *2018 second international conference on computing methodologies and communication (ICCMC)*, pages 869–873. IEEE.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Tommaso Fornaciari, Alexandra Uma, Massimo Poesio, and Dirk Hovy. 2022. Hard and soft evaluation of NLP models with BOOtSTrap SAmpling - BooStSa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–134, Dublin, Ireland. Association for Computational Linguistics.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Lijimol George and P Sumathy. 2023. An integrated clustering and bert framework for improved topic modeling. *International Journal of Information Technology*, pages 1–9.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

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.

Soumyajit Gupta, Sooyong Lee, Maria De-Arteaga, and Matthew Lease. 2023. Same same, but different: Conditional multi-task learning for demographic-specific toxicity detection. In *Proceedings of the ACM Web Conference 2023*, pages 3689–3700.

Vivek Gupta, Ankit Saw, Pegah Nokhiz, Harshit Gupta, and Partha Talukdar. 2019. Improving document classification with multi-sense embeddings. *arXiv preprint arXiv:1911.07918*.

Danula Hettiachchi, Mike Schaekermann, Tristan J McKinney, and Matthew Lease. 2021. The challenge of variable effort crowdsourcing and how visible gold can help. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–26.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. 2018. Learning word embeddings: Unsupervised methods for

fixed-size representations of variable-length speech segments. In *Interspeech 2018*. ISCA.

Soodeh Hosseini and Zahra Asghari Varzaneh. 2022. Deep text clustering using stacked autoencoder. *Multimedia Tools and Applications*, 81(8):10861–10881.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Ali Idrus. 2022. Distance analysis measuring for clustering using k-means and davies bouldin index algorithm. *TEM Journal*, 11(4):1871–1876.

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

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Ismo Kärkkäinen and Pasi Fränti. 2000. Minimization of the value of davies-bouldin index. In *Proceedings of the IASTED International Conference on Signal Processing and Communications (SPC'2000). IASTED/ACTA Press*, pages 426–432.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Attila Lengyel and Zoltán Botta-Dukát. 2019. Silhouette width using generalized mean—a flexible method for assessing clustering efficiency. *Ecology and evolution*, 9(23):13231–13243.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 299. NIH Public Access.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. *arXiv preprint arXiv:2109.04740*.

Junaid Rashid, Syed Muhammad Adnan Shah, and Aun Irtaza. 2020. An efficient topic modeling approach for text mining and information retrieval through k-means clustering. *Mehran University Research Journal of Engineering & Technology*, 39(1):213–222.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2420–2433.

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

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. *arXiv preprint arXiv:1904.01617*.

Xavier Sevillano, Germán Cobo, Francesc Alías, and Joan Claudi Socoró. 2007. Text clustering on latent thematic spaces: Variants, strengths and weaknesses. In *International Conference on Independent Component Analysis and Signal Separation*, pages 794–801. Springer.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Wei Song and Soon Cheol Park. 2007. A novel document clustering model based on latent semantic analysis. In *Third International Conference on Semantics, Knowledge and Grid (SKG 2007)*, pages 539–542. IEEE.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021a. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021b. Mbic–a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.11910*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Michael Sullivan, Mohammed Yasin, and Cassandra L Jacobs. 2023. University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

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

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. ilab at semeval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? *arXiv preprint arXiv:2305.06074*.

Dongsheng Wang, Prayag Tiwari, Mohammad Shorfuzzaman, and Ingo Schmitt. 2021. Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing. *Computer Networks*, 196:108227.

Charles Welch, Jonathan K Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. *arXiv preprint arXiv:2010.02986*.

Hyperparameter	GWSD	MBIC
Pretrained Model	`google/t5-v1_1-large`	`google/t5-v1_1-large`
Decoder Depth	2	4
Decoder Heads	4	16
Dropout	0.2979	0.0979
Learning Rate	4.85×10^{-5}	3.5×10^{-5}
Warmup Steps	332	296

Table 5: Hyperparameters for the **Cross Attention** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

Hyperparameter	GWSD	MBIC
Pretrained Model	`google/t5-v1_1-large`	`google/t5-v1_1-large`
Decoder Depth	15	12
Decoder Heads	2	16
Dropout	0.188	0.0657
Learning Rate	9.27×10^{-5}	1.13×10^{-5}
Warmup Steps	296	535
Downsampling Num. Layer	1	3
Downsampling Dimension	768	192

Table 6: Hyperparameters for the **Pooled Cross Attention** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

Haoran Xu and Philipp Koehn. 2021. Cross-lingual bert contextual embedding space mapping with isotropic and isometric conditions. *arXiv preprint arXiv:2107.09186*.

A Training Details

To aid in reproducibility, we report all training details and any relevant hyperparameters.

A.1 Hyperparameters

All models were trained using a single NVIDIA A40 GPU. A total of 1080 hours were used during training of all models. For all models, we used the AdamW optimizer (Loshchilov and Hutter, 2017) during training with weight decay 0.01. We report hyperparameters for each model and dataset in Tables 5 to 10.

From small performance gains during preliminary experiments, we disable bias across all linear layers as indicated in Fig. 2. Across every model, we found that when comparing hyperparameters for both PCA and UMAP converged to the same choices. For both methods, we found that 2 components yielded the best results. Additionally, for UMAP, we found that the optimal number of neighbours were found to be between 80-100 across all models,with a minimum distance ranging from 0.8 to 1 to yield better clustering performance.

Hyperparameter	GWSD	MBIC
Pretrained Model	`google/t5-v1_1-large`	`google/t5-v1_1-large`
Decoder Depth	4	2
Decoder Heads	2	8
Dropout	0.241	0.155
Learning Rate	4.24×10^{-5}	7.67×10^{-5}
Warmup Steps	782	426

Table 7: Hyperparameters for the **Encoder-Encoder** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

Hyperparameter	GWSD	MBIC
Pretrained Model	`roberta-large`	`roberta-large`
Learning Rate	8.405×10^{-6}	1.31×10^{-4}
Warmup Steps	131	639

Table 8: Hyperparameters for the **Classifier** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

B Cluster Metrics

B.1 Dimensionality Reduction

B.2 GWSD Cluster Validity Scores

We report the GWSD internal and external validation metrics resulting from our clustering using a k-means algorithm and our various employed dimensionality reduction techniques.

Hyperparameter	GWSD	MBIC
Pretrained Model	`gpt2-large`	`gpt2-large`
Learning Rate	9.387×10^{-7}	6.443×10^{-6}
Warmup Steps	415	637

Table 9: Hyperparameters for the **Pretrained Decoder** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

Hyperparameter	GWSD	MBIC
Pretrained Model	`gpt2-large`	`gpt2-large`
Learning Rate	3.61×10^{-4}	2.91×10^{-4}
Warmup Steps	894	778

Table 10: Hyperparameters for the **Pretrained Encoder-Decoder** models on each of our chosen datasets, obtained from running running a hyperparameter sweep for 12 hours.

	Davies-Bouldin Index	Silhouette
No dim. reduction	3.655	0.073
w/ PCA	0.491	0.56
w/ UMAP	0.565	0.53

Table 11: Dimensionality reduction effect on internal validity scores

	# Clusters	Davies-Bouldin Index ↓	Silhouette ↑	Purity-Political ↑	Purity-Education ↑
GWSD					
Cross Attention					
No dim. reduction	19	5.97	0.02	0.45	0.53
w/ PCA	19	0.49	0.55	0.49	0.48
w/ UMAP	19	0.51	0.53	0.49	0.51
Pooled Cross Attention					
No dim. reduction	16	2.73	0.08	0.45	0.65
w/ PCA	19	0.52	0.55	0.59	0.57
w/ UMAP	19	0.46	0.54	0.56	0.54
Encoder-Encoder					
No dim. reduction	18	5.77	0.02	0.53	0.34
w/ PCA	19	0.51	0.53	0.40	0.47
w/ UMAP	15	0.49	0.55	0.46	0.63
Classifier Model					
No dim. reduction	19	2.10	0.17	0.53	0.46
w/ PCA	17	0.45	0.61	0.53	0.53
w/ UMAP	18	0.95	0.46	0.44	0.51
Pretrained Decoder					
No dim. reduction	19	2.83	0.09	0.61	0.47
w/ PCA	19	0.47	0.59	0.42	0.44
w/ UMAP	17	0.49	0.55	0.49	0.51
Pretrained Encoder-Decoder					
No dim. reduction	19	2.53	0.06	0.48	0.55
w/ PCA	19	0.51	0.53	0.47	0.44
w/ UMAP	17	0.49	0.55	0.43	0.58

Table 12: Overall performance through internal (Davies-Bouldin Index, Silhouette) and external (Purity Political / Education) validity measures for the clustering component of our framework. Intra-cluster separation indicative of better overall clustering performance indicated by higher Silhouette and lower Davies-Bouldin scores. External validity, measured via inter-cluster purity, indicated by higher purity scores