Will Annotators Disagree? Identifying Subjectivity in Value-Laden Arguments

Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M. Jonker, and **Luciano C. Siebert**
Delft University of Technology, the Netherlands
{a.homayounirad,e.liscio,t.wang-12,
c.m.jonker,l.cavalcantesiebert}@tudelft.nl

Abstract

Aggregating multiple annotations into a single ground truth label may hide valuable insights into annotator disagreement, particularly in tasks where subjectivity plays a crucial role. In this work, we explore methods for identifying subjectivity in recognizing the human values that motivate arguments. We evaluate two main approaches: inferring subjectivity through value prediction vs. directly identifying subjectivity. Our experiments show that direct subjectivity identification significantly improves the model performance of flagging subjective arguments. Furthermore, combining contrastive loss with binary cross-entropy loss does not improve performance but reduces the dependency on per-label subjectivity. Our proposed methods can help identify arguments that individuals may interpret differently, fostering a more nuanced annotation process[1].

1 Introduction

Human values, spanning concepts such as benevolence and self-determination, are the motivations that guide our choice and action, and are ordered by importance to form a system of value priorities (Schwartz, 1994). Value-laden arguments are statements grounded in our personal values, which we use to motivate our choices (Bench-Capon, 2003). The identification of the values that support our arguments can reveal our deepest motivations, and as such has been recently investigated in the NLP community (Kiesel et al., 2023, 2024).

Supervised NLP methods have been proposed to identify the values that support a text segment (Kiesel et al., 2022; Liscio et al., 2022). Typically, the ground truth labels are chosen through majority aggregation of the annotations (Hoover et al., 2020; Kiesel et al., 2023) or the annotators engage in discussions to reach an agreement on the annotation (Liscio et al., 2021; Lei et al., 2024). However, due to the subjective nature of valuing (Mackie, 1988; Stroud, 1988), disagreement in the interpretation of the values that support an argument is natural. For example, consider the following argument in favor of a multi-party political system: "(it) would bring many new and fresh ideas into the forefront". Alice may associate the argument with the value of universalism since a multi-party system can provide all people with equal opportunities. Bob, instead, may connect it to the values of achievement and personal security, because a multi-party system can more effectively address issues than a single-party system and thus provide more security to its citizens[2].

Approaches relying on consensus or majority aggregation might obscure the inherent subjectivity of the identification of the values behind arguments, leading to misinterpretation or, at worst, promotion of biases stemming from the annotation process. Identifying subjectivity in value-laden arguments thus has broader implications, particularly in contexts that demand participatory deliberation and collective decision-making.

The assumption that annotations should be aggregated into a unique label is being questioned within the NLP community (Röttger et al., 2021; Weerasooriya et al., 2023; van der Meer et al., 2024), in applications ranging from hate speech detection (Kocoń et al., 2021; Mostafazadeh Davani et al., 2022) to sentiment and emotion detection (Deng et al., 2023). In line with these works, in this paper we propose methods to identify subjective value-laden arguments—that is, detect arguments where annotators may have different interpretations of the values that support them.

We envision two primary applications of this

[1] Our code is publicly available at https://github.com/Amir-Homayouni/subjectivity-value

[2] The argument and the two annotations are sourced from the ValueEval'23 Shared Task (Kiesel et al., 2023).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15237–15252
November 4-9, 2025

work. (1) Identifying arguments that might be subjective can support a more nuanced annotation process by prompting the collection of additional annotations. (2) Identifying the arguments that might lead to misinterpretation and divergent views during participatory deliberations can prompt moderators to ask additional questions that could promote self or inter-participant reflection.

Contribution We propose two different approaches to detecting subjectivity in value identification within discourse: three methods for inferring subjectivity from individual value annotations, and three methods for directly identifying subjectivity, which we further enhance with contrastive learning strategies. We validate the approaches on the Touché23-ValueEval dataset (Kiesel et al., 2023). Our results show that we can identify subjectivity in value-laden arguments, and that directly identifying subjectivity—rather than inferring it through value prediction—greatly improves performance. Additionally, leveraging contrastive loss does not improve subjectivity prediction performance but brings other advantages.

2 Related works

We review related works on identifying values in text and on subjectivity in NLP applications.

2.1 Identifying Values in Text

Identifying the value(s) that support a natural language statement has been approached through word count and sentence embedding similarity to dictionaries of value-laden words (Araque et al., 2020). More recent approaches employ supervised machine learning on annotated datasets (Alshomary et al., 2022; Huang et al., 2022; Park et al., 2024; Liscio et al., 2025; Senthilkumar et al., 2024). In particular, Kiesel et al. (2022) focuses on identifying and classifying the values underlying arguments. They successfully fine-tuned a BERT model (Devlin et al., 2019) on multi-cultural arguments, and later extended the dataset for the ValueEval challenge at SemEval'23 (Kiesel et al., 2023).

In this study, unlike the previous work that uses golden ground truths to identify values, we utilize a dataset consisting of annotation of values to arguments to explore the subjectivity of annotations in recognizing the values behind arguments.

2.2 Identifying Subjectivity

Subjectivity is playing an increasingly central role in various NLP tasks (Plank, 2022). Datasets that report individual-level annotation (e.g., (Aroyo et al., 2023)) facilitate the modeling of individual and group annotation behavior, with different annotation paradigms shown to have a great impact on data quality and model performance Röttger et al. (2021). This information allows to represent ground truth as a label distribution, preserving diverse human judgments and minority opinions Weerasooriya et al. (2023). Different methods account for subjectivity by combining annotator and annotation embeddings (Deng et al., 2023), modeling multi-annotator architectures (Mostafazadeh Davani et al., 2022), and capturing annotators' perspectives by combining their demographic information and their opinions on online content (Fleisig et al., 2023). Demographic information of annotators has been employed as a feature, however with mixed results (Goyal et al., 2022; Wan et al., 2023; Orlikowski et al., 2023). Subjectivity has also been explored as part of the sampling strategy of Active Learning methods, e.g. to select the next sample to be annotated (Baumler et al., 2023; Wang and Plank, 2023) or the next annotator that should annotate it (van der Meer et al., 2024).

These approaches primarily focus on predicting the opinions of individuals or groups of annotators. Instead, our methods identify whether we can expect disagreement among annotators identifying the values that support an argument, thus not being reliant on specific individuals or groups.

2.3 Measuring disagreement on moral ambiguity

In morally ambiguous scenarios, disagreements naturally arise due to the inherent diversity and pluralism of values people hold. Grounding theories like value pluralism suggest that such disagreements reflect legitimate differences in moral frameworks rather than mere errors (Kekes, 1996). Among others, the Moral Foundations Theory illustrates how different moral intuitions (care, fairness, authority, loyalty, sanctity) lead to divergent judgments among groups such as liberals and conservatives, making some conflicts particularly intractable (Haidt and Graham, 2007; Graham et al., 2013). Social psychologists further emphasize cognitive biases, such as naïve realism, where indi-

viduals perceive their views as objectively correct and opposing views as biased or misinformed, intensifying disagreements (Ross and Ward, 2013; Lord et al., 1979). Empirical methods from social sciences have operationalized these disagreements through inter-coder reliability metrics (e.g., Cohen's kappa, Krippendorff's alpha), recognizing that systematic variance among annotators can reflect meaningful differences in interpretation rather than random error (Krippendorff, 2004). Theories of deliberative democracy, notably by Gutmann and Thompson (2004); Rawls (2002), provide frameworks for managing moral disagreements through mutual respect and reason-giving rather than forcing consensus.

3 Methods

We propose two approaches to identify subjective value-laden arguments—that is, detect whether we expect annotators to disagree on the value annotation of a piece of text. In the first approach, Inferred Subjectivity identification (IS, Section 3.2), we train models to predict value labels for individual annotators and infer subjectivity from the variations in labels predicted across annotators. In the second approach, Direct Subjectivity identification (DS, Section 3.3), we train models to directly classify whether a given argument is subjective.

3.1 Task Formalization

Consider a dataset D composed of annotated triples (x_i, y_{ij}, a_j), where x_i is a piece of text containing an argument, and y_{ij} is the annotation of annotator a_j. Annotators can assign multiple values to each text x_i, chosen from a list of values $v_k \in V$. We aim to create a model $f(x) \in (0, 1)$ that predicts whether we expect disagreement (0) or agreement (1) among the annotators when annotating the value(s) that support an argument. We consider an argument to be subjective if at least one annotator from a group assigns a different set of values to that text compared to others.

3.2 Inferred Subjectivity Identification (IS)

In this approach, we explore three multi-annotator architectures to assess subjective value prediction, as displayed in Figure 1. They all pass the input arguments into sentence embedding and classification heads in different ways to predict multi-label values for each annotator. Appendix B presents the results of each method for value prediction. We then classify x_i as subjective for each v_k where individual annotators' predictions differ and report the subjectivity classification results.

3.2.1 IS-each: A Dedicated Model for Each Annotator

We train a multi-label classification model for each annotator a_j to predict value values based on the annotations they provided. Although straightforward, this approach is computationally expensive as it requires training a separate model for each annotator.

3.2.2 IS-shared: a Shared Model with a Dedicated Head for Each Annotator

We train a single model for all annotators, thus with shared embeddings but a different multi-label classification head for each annotator. This approach reduces computational complexity by sharing common embeddings across all annotators.

3.2.3 IS-single: One Model for All Annotators

In this approach, we train a single model for all annotators, incorporating a unique annotator identifier as part of the input. The input to the model is modified by concatenating the annotator ID (a_j) with the textual data. This method may address some of the computational constraints associated with the previous methods, however, it may not capture annotator nuances by just using annotator ID as input and their annotation as output.

3.3 Direct Subjectivity Identification (DS)

DS directly trains the model to assess whether we expect annotators to disagree in the annotation of v_k in text x_i, as displayed in Figure 2. We frame the task as a binary classification problem and compare three methods. In the first, we employ a dedicated model for each value label to predict subjectivity. In the next two, we insert a contrastive learning objective in the model, in a supervised and unsupervised manner, respectively.

3.3.1 DS-simple: A Model to Predict Subjectivity for Each Value Label

Analogously to section 3.2.1, we train a model to predict whether we expect the annotation to be subjective, for each value (v_k), thus resulting in $|V|$ binary prediction models.

3.3.2 Contrastive Loss Primer

Cross-entropy loss has been shown to have several shortcomings, such as not explicitly encouraging

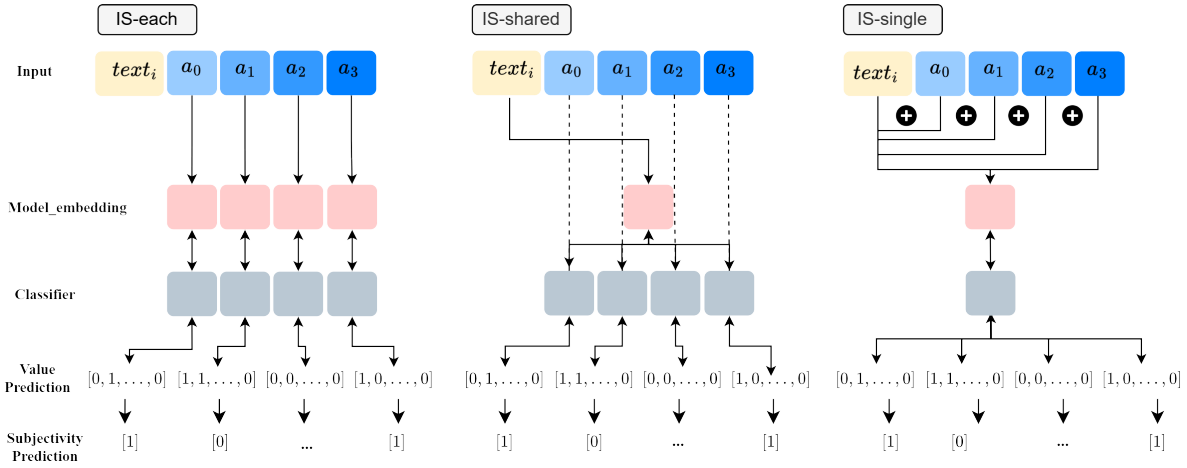


Figure 1: Representation of how the method processes input text to predict values for individual annotators and then subsequently infer subjectivity for each value label

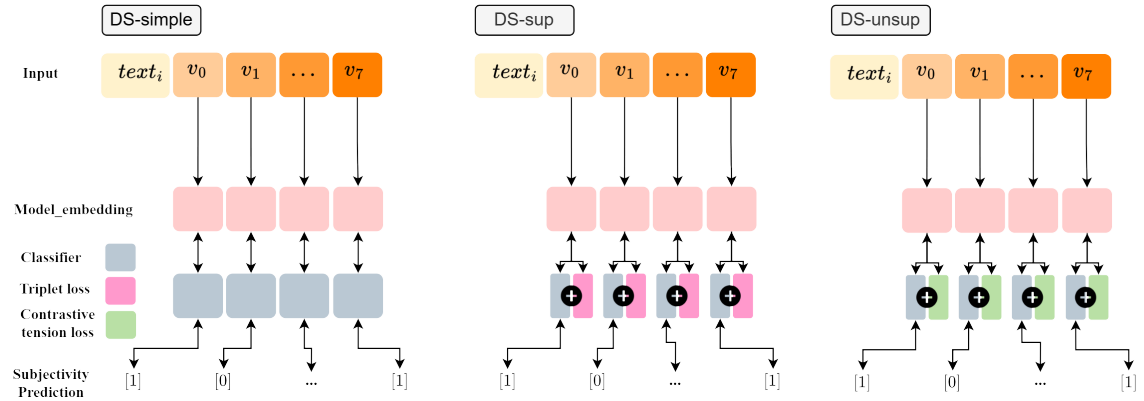


Figure 2: Representation of how the method processes input text to identify subjectivity for each value label

discriminative learning, leading to poor generalization performance (Liu et al., 2016; Cao et al., 2019). Contrastive learning as an auxiliary training objective during fine-tuning has been shown to mitigate these problems (Gunel et al., 2020), even in an unsupervised manner (Kim et al., 2021). Contrastive learning focuses on enhancing the representation of sentence embeddings by bringing semantically similar examples closer together and separating dissimilar examples (Hadsell et al., 2006).

For the following two methods, we consider a primary task as training a binary classifier to predict the subjectivity of the input text using Binary Cross-Entropy (BCE) loss, and as an auxiliary task use a Contrastive Learning (CL) loss:

$$L = L_{BCE} + \lambda L_{CL}$$

where L_{BCE} represents the binary cross-entropy loss, L_{CL} denotes the contrastive learning loss, and λ controls the relative importance of the CL loss compared to the BCE loss. The methods differ in how L_{CL} is calculated—in a supervised and unsupervised fashion, respectively.

3.3.3 DS-sup: DS-simple + Supervised Contrastive Loss

In this method, we use triplet loss ($L_{triplet}$) (Schroff et al., 2015), a supervised version of contrastive loss that exploits data labels to refine the embedding space such that examples labeled with the same class (subjective) are pushed closer together and examples with a different class (non-subjective) are pushed further apart.

Every training sample is composed of a triple (x_A, x_P, x_N), where x_A is the anchor sample, x_P (positive sample) is a randomly selected sample with the same label as x_A, and x_N (negative sample) is a randomly selected sample with a different label from x_A. $L_{triplet}$ is then defined as:

$$L_{triplet} = \max\{d(z_A, z_P) - d(z_A, z_N) + m,\ 0\}$$

where z_A, z_P, and z_N are the normalized representations in embedding space of the anchor, positive,

and negative inputs, respectively. $m > 0$ is the margin hyperparameter that enforces a minimum separation between positive and negative pairs.

3.3.4 DS-unsup: DS-simple + Unsupervised Contrastive Loss

In this method, we employ contrastive tension loss ($L_{tension}$) (Carlsson et al., 2021), an unsupervised contrastive loss. Whereas the previous method employed label information to shape the embedding space, this method instead ensures that semantically similar samples are pushed closer together while dissimilar examples are pushed further apart. $L_{tension}$ is then defined as:

$$L_{\text{tension}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\frac{\text{sim}\left(z_i,z_{i+}\right)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\text{sim}\left(z_i,z_j\right)}{\tau}\right)}\right),$$

where z_i is the embedding of the *anchor* sample, z_{i+} is the embedding of a *positive* sample , z_j are all samples in the batch (acting as negatives when $j \neq i^+$), $\text{sim}(\cdot,\cdot)$ is a similarity function, $\tau > 0$ is a temperature hyperparameter, and N is the batch size.

4 Experimental Setup

We describe the models, dataset, and evaluation metrics for our experiments.

4.1 Models

We test all proposed methods with BERT-base (Devlin et al., 2019) as model embedding. Next, we fine-tune the Llama-3.1-8B-Instruct model (Touvron et al., 2023) for two variants of the two approaches (IS-single and DS-simple), as further elaborated in Section 5. We compare proposed methods against a baseline that randomly predicts subjectivity. While our experiments are only conducted with two models, we propose a model-agnostic approach that is not limited to these models. Appendix A.3 details the used hyperparameters.

4.2 Dataset

We use the Touché23-ValueEval dataset (Mirzakhmedova et al., 2023), which, to the best of our knowledge, is the only dataset available that includes value annotations from multiple annotators for each instance. The dataset is composed of 9324 natural language arguments annotated with a taxonomy of 54 values (multi-label annotation) derived from the Schwartz Value Survey (Schwartz et al., 2012) and distributed in four hierarchical levels. Argument datasets are almost exclusively from a Western background on controversial topics namely religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms. We utilize annotations representing the crowd workers' original annotations (before being aggregated into a single ground truth label) that have all been carried out by annotators from a Western background. To effectively demonstrate our methods while minimizing the computational load, we selected the eight most frequently annotated values from level 2. These values were annotated by the four annotators who, among the 39 annotators, had the highest annotation overlap. The selected values are Achievement (`Ach`), Power: resources (`Pow`), Security: personal (`Sec-p`), Security: societal (`Sec-s`), Conformity: rules (`Con`), Benevolence: caring (`Ben-car`), Benevolence: dependability (`Ben-dep`), Universalism: concern (`Uni`). Figure 3 presents the distribution between the subjective and non-subjective annotations for each of the selected values, and Appendix A.1 provides more information on the dataset, annotators, and selected values.

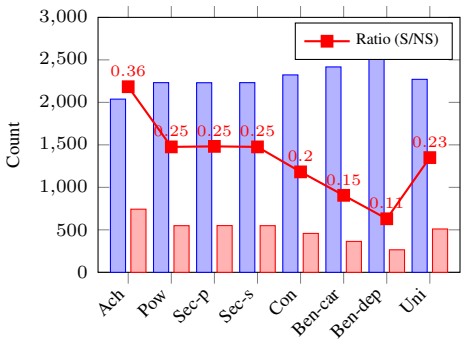


Figure 3: Counts and ratio of subjective (S, red) and non-subjective (NS, blue) labels per each selected value.

Furthermore, for the DS approaches, we augment the minority class to balance the dataset by paraphrasing the minority class sentences (Alisetti, 2021) (see Appendix A.2 for additional details). Data augmentation is not possible in the IS approach due to the multi-label nature of the task—augmenting data for one label would also impact the other labels annotated on the same sentence.

4.3 Evaluation Metrics

We report precision, recall, and F_1-score on the test set (which is fixed, as detailed in Appendix A.1), per value and averaged over the values. In addition, we report the Spearman correlation (ρ) between the per-value F_1-scores and the subjective to non-subjective ratios (Figure 3) to investigate the correlation between subjectivity level and performance.

5 Results

Table 1 reports the performance of the different approaches averaged over the selected values, whereas Table 2 reports the per-value performance.

Method	P	R	F_1	ρ
IS-each (BERT)	0.34	0.40	0.36	0.88
IS-shared (BERT)	0.31	0.22	0.25	0.85
IS-single (BERT)	0.35	0.30	0.32	0.86
IS-single (Llama)	0.40	0.26	0.30	0.88
DS-simple (BERT)	0.85	0.61	0.70	-0.60
DS-sup (BERT)	0.85	0.62	0.71	-0.32
DS-unsup (BERT)	0.70	0.61	0.65	0.24
DS-simple (Llama)	0.84	0.76	0.80	-0.97
Baseline (random)	0.53	0.50	0.51	-0.12

Table 1: Average precision (P), recall (R), and F_1-score across the selected value labels, together with the correlation (ρ) between F_1-scores and subjectivity ratio.

5.1 Comparison across Methods

First, we compare the two primary approaches—inferring subjectivity from value prediction (IS) versus direct subjectivity identification (DS). We notice that the latter demonstrates superior performance and that even the random baseline mostly outperforms the IS approach (except for IS-each for recall in Ach, Pow, and Sec-s). This may indicate that predicting value solely through annotation may not be sufficient to capture individual subjective preferences and infer subjectivity. However, the DS approach consistently outperforms the baseline in all metrics.

Second, we compare the methods that infer subjectivity from value prediction (IS). We observe that IS-each consistently outperforms the others. This is presumably because having a dedicated model for each annotator better captures their annotation tendencies when compared to having a shared embedding layer (IS-shared) or having a fully shared model (IS-single). Moreover, IS-single consistently outperforms IS-shared, which shows that differentiating the input text with an annotator ID shows a better performance compared to having a dedicated head for each annotator.

Finally, we compare the approaches that directly infer subjectivity (DS). We notice that DS-simple and DS-sup generally perform similarly and better than DS-unsup. However, we observe that for more subjective values (such as Ach) the performances with the three methods are comparable, whereas for less subjective values (such as Ben-dep) the results of DS-unsup are significantly worse. This is also supported by the correlation scores in Table 1, which show a moderate positive correlation between the DS-unsup results and per-value subjectivity.

5.2 Comparison across Values

Next, we compare the results across values. Table 1 shows that the performances with the IS approach are consistently correlated with the subjectivity of the value annotations. With the DS approach, instead, only the DS-unsup results are correlated with subjectivity—that is, for the best-performing methods (DS-simple and DS-sup), lower subjectivity leads to better performances.

We conjecture that the correlation is strongly positive for all IS methods because a more balanced class distribution leads to better value prediction, which in turn leads to better subjectivity identification. Instead, for the DS methods, which directly identify subjectivity, we observe that precision is consistently higher than recall, exceptions are DS-simple (Llama) for `Ach` and DS-unsup, DS-simple (Llama) for `Sec-p`. This is likely negatively correlated with subjectivity because the model may overfit to simpler examples hurting performance on more subtle (high subjective) values.

Finally, DS-simple has a stronger negative correlation with per-value subjectivity than DS-sup, suggesting it systematically achieves a higher F_1-score on values for which fewer annotators disagree. On the other hand, DS-sup is more balanced across value subjectivity. We conjecture this is due to the fact that supervised contrastive loss pushes the embeddings for the same label closer together, regardless of the value's overall subjective level. This shapes the embedding space more consistently across both highly subjective and less subjective values, resulting in a mitigation of overfitting.

Method	Achievement			Power: resources			Security: personal			Security: societal		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
IS-each (BERT)	0.40	0.56	0.46	0.41	0.50	0.45	0.35	0.44	0.39	0.36	0.50	0.42
IS-shared (BERT)	0.37	0.29	0.32	0.38	0.26	0.31	0.32	0.20	0.24	0.41	0.34	0.37
IS-single (BERT)	0.38	0.39	0.38	0.40	0.33	0.36	0.37	0.40	0.38	0.38	0.36	0.37
IS-single (Llama)	0.41	0.39	0.40	0.44	0.29	0.35	0.46	0.41	0.33	0.45	0.31	0.37
DS-simple (BERT)	**0.81**	0.57	0.67	0.82	0.61	0.69	**0.83**	0.66	0.73	**0.89**	0.58	0.70
DS-sup (BERT)	0.78	0.62	0.69	0.84	0.58	0.69	**0.83**	0.64	0.72	0.84	0.65	0.73
DS-unsup (BERT)	0.65	0.60	0.62	0.73	0.55	0.63	0.65	0.72	0.68	0.71	0.67	0.68
DS-simple (Llama)	0.71	**0.76**	**0.74**	**0.94**	**0.65**	**0.77**	0.73	**0.80**	**0.76**	0.82	**0.74**	**0.78**
Baseline (random)	0.55	0.49	0.51	0.53	0.50	0.52	0.51	0.48	0.49	0.55	0.49	0.52

	Conformity: rules			Benevolence: caring			Benevolence: depend.			Universalism: concern		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
IS-each (BERT)	0.30	0.35	0.33	0.33	0.42	0.37	0.20	0.24	0.22	0.36	0.2	0.24
IS-shared (BERT)	0.28	0.18	0.22	0.28	0.18	0.22	0.14	0.15	0.14	0.32	0.16	0.20
IS-single (BERT)	0.30	0.26	0.27	0.35	0.27	0.30	0.23	0.19	0.21	0.37	0.20	0.25
IS-single (Llama)	0.42	0.19	0.26	0.35	0.22	0.27	0.33	0.13	0.18	0.35	0.17	0.23
DS-simple (BERT)	**0.89**	0.60	0.71	0.86	0.62	0.71	**0.90**	0.60	0.72	0.85	0.59	0.70
DS-sup (BERT)	**0.89**	0.61	0.72	**0.89**	0.56	0.69	0.86	0.67	0.75	0.85	0.59	0.69
DS-unsup (BERT)	0.70	0.68	0.69	0.78	0.52	0.62	0.66	0.55	0.60	0.70	0.59	0.64
DS-simple (Llama)	0.88	**0.75**	**0.81**	0.88	**0.78**	**0.83**	0.88	**0.86**	**0.87**	**0.91**	**0.71**	**0.80**
Baseline (random)	0.50	0.50	0.50	0.52	0.50	0.51	0.52	0.52	0.52	0.53	0.50	0.52

Table 2: Precision (P), recall (R), and F_1-score of the subjectivity prediction per value. In bold, highlight the best-performing method and underline the second-best-performing method for each value and metric.

5.3 Comparison across Models

Finally, we compare the results across models. For the IS approach, we decided to test the Llama model with the IS-single method to investigate whether using a more powerful model can compensate for the difference between using a single model for all annotators (IS-single) or one model per annotator (IS-each). However, we observe no improvement over the IS-single results with the BERT model. We conjecture that this is due to the structure of IS-single—that is, adding an annotator ID to each argument to differentiate between annotators. Such information is evidently not sufficient to differentiate across annotators.

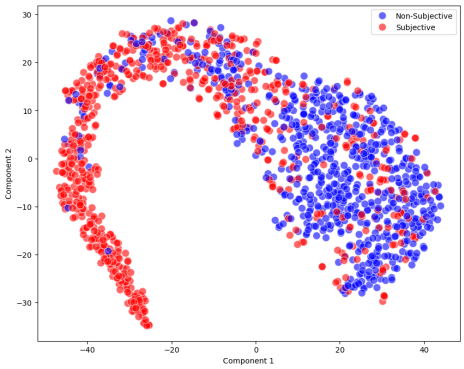
Next, given the comparable performances between DS-simple and DS-sup, we decide to train Llama with DS-simple for simplicity. Differently from the IS approach, in the DS approach, we observe a performance improvement over BERT, particularly in recall across all values and in precision for values `Pow`, and `Uni`, in line with the difference in state-of-the-art between the two models.

6 Discussion

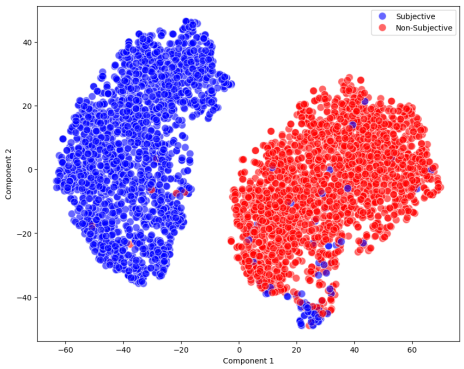
In this work, we explored two distinct approaches to address the challenge of identifying subjectivity in value-laden arguments. We discuss our results across methods and value labels.

6.1 Comparison across Methods

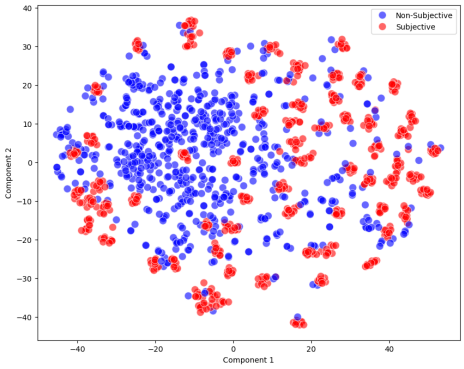
DS is better, but IS also has its merits. The superior performance of the DS approach can be likely attributed to its focused objective, which simplifies the model's learning process by concentrating on distinguishing between subjective and non-subjective instances. This targeted focus allows the model to capture patterns indicative of subjectivity, leading to improved recall and F_1-score rates. While it is intuitive that direct subjectivity identification might outperform inference-based methods, we argue that inferring subjectivity from value predictions remains valuable in scenarios where understanding individual value preferences provides meaningful insights. For instance, exploring why an argument is subjective by looking at each annotator's value prediction and understanding their different interpretation. However, improving subjectivity identification through value prediction may require the development of methodologies that explicitly model the relationship between value prediction and subjectivity identification. For example, incorporating individual subjective preferences derived not only from value annotations but also from additional factors such as annotators' backgrounds can enhance this link.



(a) BERT-base



(b) BERT embeddings trained with DS-sup.



(c) BERT embeddings trained with DS-unsup.

Figure 4: Comparison of 2D sentence embeddings visualizations obtained through t-SNE. (Top) Embeddings from the original BERT. (Middle) Embeddings from the BERT instance fine-tuned with DS-sup. (Bottom) Embeddings from the BERT instance fine-tuned with DS-unsup. Red dots correspond to subjective and blue to non-subjective.

Moreover, the subjectivity of value-laden arguments may vary across different contexts and demographic groups (Liscio et al., 2021). The second approach, which directly predicts subjectivity irrespective of individual annotators, offers a pathway for active learning strategies to update and fine-tune the model for new contexts and diverse populations.

Contrastive loss brings additional advantages. Despite the comparable subjectivity prediction performance, optimizing binary cross-entropy loss with a contrastive learning objective makes the resulting embeddings more suitable for calculating similarities between samples. We confirm this by visualizing the BERT embedding space for value `Ben-dep` resulting from the three DS methods, in Figure 4 (similar patterns were observed for all considered values). DS-sup pushes the embeddings to have a better separation between subjective and non-subjective classes, while DS-unsup groups semantically similar samples.

Improved embeddings can be leveraged to support the annotation process during an active learning procedure. For instance, we can use the DS-sup embeddings to present annotators with instances that lie near the decision boundary between subjective and non-subjective classes, enabling annotators to reflect on samples that highlight areas of uncertainty. Additionally, utilizing the DS-unsup to retrieve semantically similar samples that are identified differently in terms of subjectivity, provides annotators with contextualized reference points. This approach may support a more nuanced annotation process by allowing reflection on samples where human judgment is most needed. Moreover, this strategy can be extended beyond annotation to support participatory democracy, equipping deliberation moderators with useful information to promote reflection and facilitate more thoughtful discussions among participants.

6.2 Comparison across Values

Our findings are supported by Schwartz (1994), who asserts that values like Achievement and Power are more closely tied to personal interests, whereas Universalism and Conformity are associated with broader societal concerns and the welfare of others. Security and Conformity are boundary values. They are primarily concerned with others' interests, but their goals also regulate the pursuit of their own interests. Hence, due to the individualized nature of some values which are tied to personal experiences and individual goals such as Achievement tend to exhibit higher subjectivity compared to Universalism which is grounded in broader ethical principles that are more widely shared across different cultures and societies.

7 Conclusion

We introduce multiple approaches to identify subjectivity in value-laden arguments. Applying our methods to the Touché23-ValueEval dataset, we demonstrated that directly identifying subjectivity, as opposed to inferring it through value prediction, significantly enhances performance. Implementing a dual-task strategy that combines contrastive loss with BCE loss does not directly improve subjectivity identification, but leads to a model that is less dependent on per-label subjectivity. Finally, using a state-of-the-art model improves performance for direct subjectivity identification, but not for inferred subjectivity identification.

We envision a combination of our proposed approaches as future work, e.g. by combining value and subjectivity prediction to capture individual annotators' perspectives in combination with their annotations, thereby potentially improving the identification of subjectivity in value-laden arguments. Datasets encompassing diverse perspectives, including varying demographics, lived experiences, and moral values (Waseem, 2016; Patton et al., 2019), such as the recently introduced D3CODE dataset (Mostafazadeh Davani et al., 2024), facilitate the integration of these two aspects.

Finally, we acknowledge that differences in annotation (i.e., subjectivity) can be confounded with noise in the annotation process. Previous work addressed this issue by facilitating deliberation among crowd workers (Schaekermann et al., 2018) or by assessing the validity of the explanations provided by annotators for their responses (Weber-Genzel et al., 2024). Our work can be instrumental in identifying potentially subjective annotations to support such approaches.

8 Limitations

Our evaluation is confined to the Touché23-ValueEval dataset (Mirzakhmedova et al., 2023), primarily due to the scarcity of datasets within the value community that include annotator-level annotations. This limitation underscores the importance of developing datasets that collect a diverse set of annotators' value annotations. Such datasets would facilitate a more robust evaluation of methods aimed at modeling subjectivity.

The first two methods from ISV including —ISV-each, ISV-shared, and DS-simple are computationally expensive due to their reliance on dedicated embeddings and classifiers. Although these methods provide insights into annotator-specific tendencies, their practicality in large-scale applications is limited. Future research could focus on optimizing these models for scalability or developing lightweight alternatives. One approach for optimizing ISV is to merge annotators who share the same perspective into similar groups and apply an active learning strategy to find the most diverse and useful information to infer subjectivity. The same approach can also be applied for DS-simple to only train the data with the most informative information in terms of subjectivity detection, as Ds-unsup which semantically improves embedding can be utilized for this.

Our study focuses exclusively on classification tasks, which limits the generalizability of our findings to other NLP tasks, such as summarization and question-answering, and to fields beyond NLP, such as Reinforcement Learning and Inverse Reinforcement Learning. Investigating how subjectivity manifests in these tasks and domains remains an important avenue for future work

To use the best of our proposed methods to identify the sources of disagreement and promote reflection, our annotators must be diverse and represent different ranges of moral values, beliefs, and backgrounds. This diversity should align with the perspectives of individuals involved in the deliberation process, ensuring that annotators reflect a broad spectrum of participants. The dataset we use in this study has no intention to be used for a specific deliberation setting, and we had no control over the notion of alignment between annotators' subjective viewpoints and participants in deliberation. Hence, as values are context-specific Vargo and Lusch (2015); Horbel et al. (2016); Chandler and Vargo (2011); Edvardsson et al. (2010); Liscio et al. (2021), to be able to identify subjective value-laden arguments more meaningful with respect to the new deliberation setting, we encourage users of this method to either train the methods in a new context or fine-tune the model with representative annotators of deliberation.

Acknowledgments

This project is part of the AiBLE lab, which receives support from the TU Delft AI Labs program. This work was partly funded by AlgoSoc, ESDiT, and the Hybrid Intelligence Centre (project numbers 024.005.017, 024.004.031, and 024.004.022). Any opinions, findings, and conclusions or recom-

mendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OCW or those of the AlgoSoc consortium as a whole. GitHub Copilot was used as an assistant programming tool.

References

Sai Vamsi Alisetti. 2021. Paraphrase generator with t5.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher Michael Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *ArXiv*, abs/2306.11247.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Trevor J. M. Bench-Capon. 2003. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Neural Information Processing Systems*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

J. D. Chandler and S. L. Vargo. 2011. Contextualization and value-in-context: how context frames exchange. *Marketing Theory*, 11:35–49.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

B. Edvardsson, B. Tronvoll, and T. Gruber. 2010. Expanding understanding of service exchange and value co-creation: a social construction approach. *Journal of the Academy of Marketing Science*, 39:327–339.

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, Singapore. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Preprint*, arXiv:2205.00501.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *ArXiv*, abs/2011.01403.

Amy Gutmann and Dennis F Thompson. 2004. *Why deliberative democracy?* Princeton University Press.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

C. Horbel, B. Popp, H. Woratschek, and B. Wilson. 2016. How context shapes value co-creation: spectator experience of sport events. *The Service Industries Journal*, 36:510–531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xiaolei Huang, Alexandra S. Wormley, and Adam H. Cohen. 2022. Learning to adapt domain shifts of moral values via instance weighting. *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*.

John Kekes. 1996. The morality of pluralism.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. ACL.

Johannes Kiesel et al. 2024. Overview of Touché 2024: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. 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 2528–2540, Online. Association for Computational Linguistics.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, P. Milkowski, and Przemyslaw Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024. EMONA: Event-level moral opinions in news articles. 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 5239–5251, Mexico City, Mexico. Association for Computational Linguistics.

Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Enrico Liscio, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. 2025. Value preferences estimation and disambiguation in hybrid participatory systems. *Journal of Artificial Intelligence Research*, 82:819–850.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 799–808, Online. IFAAMAS.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*.

Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098.

John L Mackie. 1988. The subjectivity of values. *Essays on moral realism*, pages 95–118.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touché23-valueeval dataset for identifying human values behind arguments. In *International Conference on Language Resources and Evaluation*.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Jeongwoo Park, Enrico Liscio, and Pradeep Murukannaiah. 2024. Morality is non-binary: Building a pluralist moral sentence embedding space using contrastive learning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 654–673, St. Julian's, Malta. Association for Computational Linguistics.

Desmond Upton Patton, Philipp Blandfort, William R. Frey, Michael B. Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Hawaii International Conference on System Sciences*.

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.

John Rawls. 2002. *John Rawls: Political liberalism and the law of peoples*, volume 2. Taylor & Francis.

Lee Ross and Andrew Ward. 2013. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge*, pages 103–135. Psychology Press.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. In *North American Chapter of the Association for Computational Linguistics*.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Shalom H. Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103 4:663–88.

Rithik Appachi Senthilkumar, Amir Homayounirad, and Luciano Cavalcante Siebert. 2024. Leveraging large language models to identify the values behind arguments. In *International Workshop on Value Engineering in AI*, pages 87–103. Springer.

Barry Stroud. 1988. The study of human nature and the subjectivity of value. *The Tanner Lectures on Human Value*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Vamsi. T5 paraphrase paws. `https://huggingface.co/Vamsi/T5_Paraphrase_Paws`. Accessed: 2024-12-06.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

S. L. Vargo and R. F. Lusch. 2015. Institutions and axioms: an extension and update of service-dominant logic. *Journal of the Academy of Marketing Science*, 44:5–23.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: quantifying annotation disagreement using demographic information. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Xinpeng Wang and Barbara Plank. 2023. Actor: Active learning with annotator-specific classification heads to embrace human label variation. *Preprint*, arXiv:2310.14979.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

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

A Additional Experimental Details

A.1 Dataset details

we use The Touché23-ValueEval dataset by(Mirzakhmedova et al., 2023), which comprises 5270 natural language arguments that are annotated for consolidated taxonomy of 54 values (multi-class annotation). Value taxonomy is categorized on the more abstract levels 2–4 which are derived mainly from the Schwartz Value Survey (Schwartz et al., 2012). Table A1 shows the example of the original dataset. annotations have all been carried out by annotators from a Western background. This dataset is distributed under CC BY-SA 4.0. The data is split so that approximately 78% of the samples are used for training, 22% for testing, and within the training set, a further 10% is reserved for validation. The annotation process involved crowdsourcing on MTurk using a custom three-part interface designed for speed and expertise. The interface presented arguments in a scenario and asked annotators to identify relevant values by answering a yes/no question. Annotators were instructed to select one to five values per argument. The average time spent by annotators on each argument was 2 minutes and 40 seconds. For additional information regarding the dataset, please refer to (Kiesel et al., 2023)

We also use the premise, and we reference it as the argument.

Argument ID	Worker ID	Premise	Simplified_Value_lvl2_ann
A01001	W014	if entrapment can serve to more easily capture...	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
A01001	W020	if entrapment can serve to more easily capture...	[0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, ...]
A01001	W024	if entrapment can serve to more easily capture...	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
A01002	W014	we should ban human cloning as it will only ca...	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
A01002	W020	we should ban human cloning as it will only ca...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
...	...	...	...

Table A1: Data structure of argument annotation

Following Figure A1 we selected the first four annotators who annotated the most.

Based on Figure A2 we selected top 8 Most Annotated Value Categories, Category 14: 3613 annotations : Benevolence: caring(Value_5) , Category 8: 3055 annotations : Security: personal(Value_2), Category 4: 2948 annotations:

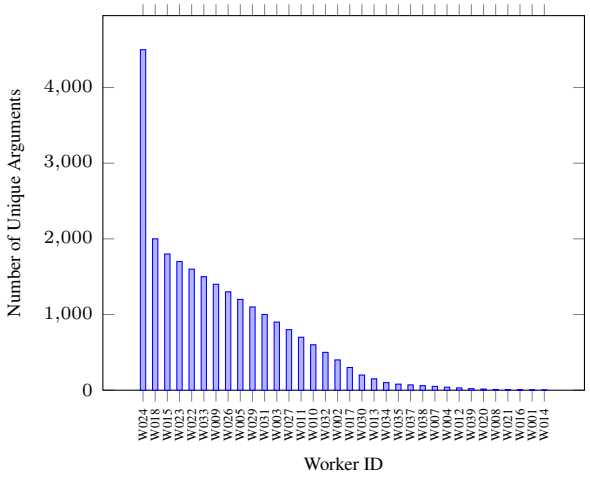


Figure A1: Number of Unique Arguments Annotated by Each Worker

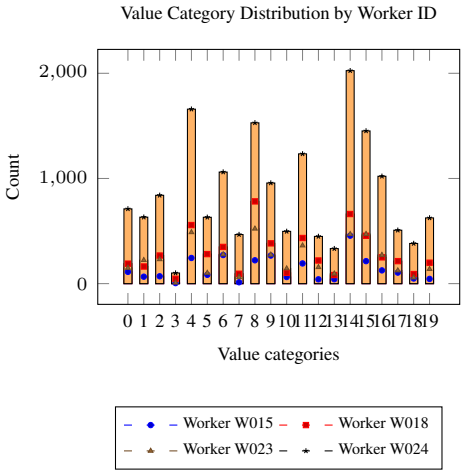


Figure A2: Value Category Distribution by Worker ID

Achievement(Value_0), Category 15: 2592 annotations : Benevolence: dependability(Value_6), Category 11: 2222 annotations : Conformity: rules(Value_4), Category 6: 1960 annotations : Power: resources(Value_1), Category 9: 1882 annotations : Security: societal(Value_3), Category 16: 1669 annotations : Universalism: concern (Value_7)

A.1.1 Fleiss Kappa score

Fleiss' kappa is a statistical metric used to evaluate the consistency of agreement among multiple raters when they assign categorical ratings to various items (Fleiss, 1971). As can be seen A3 Value_0, Value_7, Value_3 has fair agreement, and Value_1, Value_2, Value_5 has a moderate agreement and Value_6 has a substantial agreement.

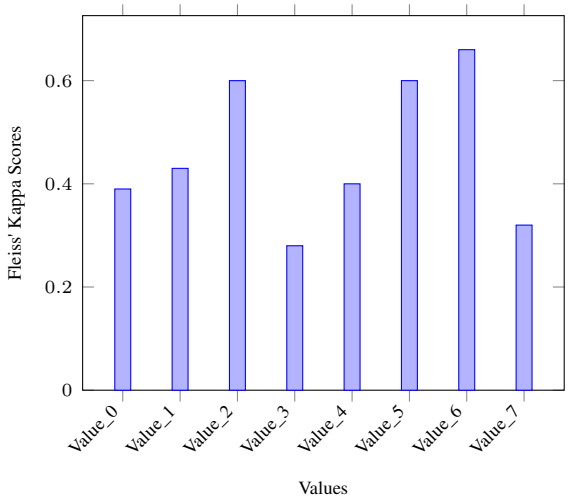


Figure A3: Fleiss' Kappa Scores per values

A.2 Paraphrasing

We utilize the Vamsi model, a T5-based transformer fine-tuned on the PAWS dataset for paraphrase generation. This model generates diverse paraphrased versions of input sentences A2 list the hyperparameter used.

Hyperparameter	Value
Sampling Method	Top-k Sampling
Temperature	2.0
Top-k	40
Top-p (Nucleus Sampling)	0.85
Repetition Penalty	1.5

Table A2: Paraphrasing

A.3 Hyperparameters and Infrastructure

For BERT, computational experiments were run on a machine containing RTX 2080 Ti GPU. For Llama, computational experiment run on NVIDIA A40 with 2 x AMD EPYC 7413 24-Core Processor. Below are the hyperparameters used for each six methods.

Originally, Llama-3.1-8B-Instruct is a causal language model designed for text generation, to adapt it for classification, we add a fully trainable classification head on top of the base model. The fully trainable classification head is optimized in conjunction with low-rank adaptation (LoRA) (Hu et al., 2021) adapters. To lower the task's computational cost, we utilize 4-bit quantization.

B Extended Results on value prediction and subjectivity prediction with STD

Hyperparameter	Value
Batch size	16
Learning rate	1e-5
Max sequence length	128
Epochs	10
Optimizer	AdamW
Pooling strategy	Mean

Table A3: Hyperparameters for DBV, SBV-ind, SBV-all

Hyperparameter	Value
Batch size	16
Learning rate	1e-5
Epochs	5
Optimizer	AdamW
Pooling strategy	Mean

Table A4: Hyperparameters for DBS

Hyperparameter	Value
Batch size	16
Learning rate	1e-5
Epochs	5
Optimizer	AdamW
Margin	1.0
Alpha (weight for triplet loss)	1.0
Pooling strategy	Mean

Table A5: Hyperparameters for DBS-SC

Hyperparameter	Value
Batch size	64
Learning rate	1e-5
Epochs	5
Optimizer	AdamW
Alpha (weight for Contrastive loss)	5.0
Pooling strategy	Mean

Table A6: Hyperparameters for DBS-UC

Hyperparameter	Value
Batch size	8
Learning rate	2e-5
Epochs	5, 1
Optimizer	AdamW
lora alpha	16
lora dropout	0.1

Table A7: Hyperparameters for IS-single (Llama), and DS-simple (Llama) with 1 epoch

Method	Label 0			Label 1			Label 2			Label 3		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
DBCV-W015	0.54±0.02	0.51±0.02	0.52±0.02	0.59±0.04	0.37±0.02	0.45±0.03	0.57±0.04	0.37±0.10	0.44±0.08	0.87±0.09	0.23±0.06	0.36±0.07
DBCV-W023	0.65±0.01	0.41±0.11	0.49±0.08	0.70±0.05	0.65±0.04	0.67±0.01	0.64±0.03	0.63±0.05	0.64±0.02	0.60±0.05	0.26±0.10	0.36±0.09
DBCV-W018	0.67±0.04	0.48±0.04	0.56±0.02	0.69±0.02	0.69±0.04	0.69±0.02	0.62±0.03	0.67±0.05	0.64±0.02	0.63±0.09	0.27±0.11	0.36±0.08
DBCV-W024	0.68±0.02	0.74±0.03	0.71±0.01	0.69±0.05	0.62±0.05	0.65±0.03	0.66±0.02	0.72±0.02	0.69±0.01	0.59±0.01	0.50±0.09	0.54±0.05
SBDCV-W015	0.52±0.03	0.63±0.09	0.57±0.03	0.48±0.12	0.47±0.16	0.45±0.04	0.37±0.08	0.66±0.13	0.46±0.05	0.42±0.10	0.38±0.15	0.38±0.10
SBDCV-W023	0.59±0.05	0.57±0.11	0.57±0.04	0.66±0.05	0.66±0.04	0.66±0.02	0.60±0.08	0.81±0.05	0.69±0.05	0.47±0.03	0.42±0.06	0.44±0.03
SBDCV-W018	0.56±0.04	0.66±0.03	0.60±0.02	0.72±0.04	0.63±0.05	0.67±0.01	0.53±0.06	0.77±0.04	0.62±0.03	0.42±0.06	0.38±0.13	0.38±0.05
SBDCV-W024	0.70±0.02	0.71±0.04	0.71±0.01	0.70±0.03	0.61±0.03	0.65±0.01	0.67±0.02	0.71±0.01	0.69±0.01	0.60±0.02	0.50±0.06	0.54±0.03
SBCV-W015	0.54±0.01	0.62±0.04	0.58±0.02	0.63±0.06	0.32±0.08	0.42±0.07	0.63±0.06	0.39±0.07	0.48±0.07	0.67±0.15	0.35±0.12	0.43±0.08
SBCV-W023	0.63±0.03	0.56±0.06	0.59±0.03	0.74±0.03	0.58±0.08	0.64±0.05	0.71±0.04	0.65±0.04	0.67±0.03	0.51±0.04	0.45±0.05	0.47±0.02
SBCV-W018	0.66±0.03	0.54±0.05	0.59±0.02	0.69±0.03	0.71±0.03	0.70±0.02	0.68±0.03	0.67±0.02	0.67±0.02	0.50±0.03	0.35±0.05	0.41±0.04
SBCV-W024	0.70±0.03	0.75±0.03	0.72±0.01	0.69±0.04	0.62±0.05	0.65±0.02	0.66±0.02	0.72±0.03	0.69±0.01	0.56±0.02	0.59±0.07	0.57±0.03
SBV-all-Llama-3.1-FT-W015	0.68	0.43	0.53	0.53	0.58	0.55	0.44	0.42	0.43	0.67	0.12	0.21
SBV-all-Llama-3.1-FT-W023	0.55	0.53	0.54	0.66	0.67	0.67	0.65	0.77	0.71	0.60	0.35	0.44
SBV-all-Llama-3.1-FT-W018	0.61	0.48	0.54	0.70	0.81	0.75	0.66	0.68	0.67	0.61	0.34	0.44
SBV-all-Llama-3.1-FT-W024	0.69	0.70	0.69	0.71	0.71	0.71	0.70	0.56	0.73	0.67	0.43	0.52

	Label 4			Label 5			Label 6			Label 7		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
DBCV-W015	0.49±0.14	0.28±0.05	0.35±0.06	0.63±0.06	0.28±0.06	0.39±0.05	0.68±0.07	0.45±0.04	0.54±0.03	0.00±0.00	0.00±0.00	0.00±0.00
DBCV-W023	0.56±0.05	0.46±0.05	0.50±0.04	0.65±0.04	0.30±0.08	0.41±0.07	0.61±0.03	0.43±0.03	0.50±0.03	0.77±0.09	0.16±0.02	0.26±0.03
DBCV-W018	0.64±0.02	0.48±0.06	0.54±0.04	0.75±0.11	0.31±0.06	0.43±0.04	0.68±0.04	0.46±0.06	0.55±0.03	0.75±0.03	0.27±0.02	0.39±0.03
DBCV-W024	0.61±0.02	0.54±0.05	0.57±0.02	0.68±0.05	0.51±0.07	0.58±0.03	0.70±0.04	0.49±0.05	0.57±0.02	0.58±0.10	0.22±0.08	0.31±0.07
SBDCV-W015	0.33±0.04	0.60±0.16	0.41±0.03	0.58±0.05	0.48±0.10	0.52±0.06	0.57±0.10	0.67±0.18	0.59±0.07	0.06±0.08	0.10±0.14	0.07±0.10
SBDCV-W023	0.50±0.06	0.56±0.05	0.52±0.02	0.51±0.05	0.48±0.08	0.49±0.02	0.52±0.05	0.54±0.07	0.53±0.03	0.56±0.19	0.32±0.13	0.37±0.03
SBDCV-W018	0.53±0.09	0.63±0.04	0.57±0.03	0.54±0.04	0.53±0.06	0.53±0.02	0.55±0.06	0.57±0.12	0.55±0.04	0.58±0.27	0.35±0.12	0.38±0.03
SBDCV-W024	0.61±0.02	0.52±0.03	0.56±0.02	0.68±0.03	0.48±0.05	0.56±0.03	0.67±0.02	0.50±0.05	0.57±0.03	0.65±0.09	0.21±0.03	0.31±0.02
SBCV-W015	0.61±0.06	0.38±0.06	0.47±0.06	0.64±0.11	0.48±0.10	0.54±0.05	0.74±0.03	0.50±0.05	0.60±0.04	0.59±0.13	0.24±0.06	0.34±0.08
SBCV-W023	0.64±0.03	0.43±0.05	0.51±0.04	0.52±0.04	0.39±0.07	0.44±0.04	0.68±0.03	0.44±0.07	0.53±0.05	0.62±0.08	0.31±0.03	0.41±0.01
SBCV-W018	0.67±0.03	0.50±0.07	0.57±0.04	0.68±0.05	0.33±0.09	0.44±0.07	0.68±0.05	0.48±0.08	0.56±0.05	0.67±0.03	0.33±0.02	0.44±0.02
SBCV-W024	0.61±0.02	0.55±0.01	0.58±0.01	0.65±0.06	0.54±0.06	0.59±0.02	0.69±0.03	0.58±0.05	0.63±0.02	0.46±0.03	0.42±0.05	0.44±0.02
SBV-all-Llama-3.1-FT-W015	0.55	0.32	0.41	0.74	0.36	0.48	0.88	0.54	0.67	0.38	0.12	0.18
SBV-all-Llama-3.1-FT-W023	0.77	0.40	0.52	0.62	0.35	0.45	0.73	0.47	0.57	0.58	0.38	0.46
SBV-all-Llama-3.1-FT-W018	0.79	0.34	0.48	0.86	0.30	0.44	0.81	0.55	0.65	0.77	0.25	0.38
SBV-all-Llama-3.1-FT-W024	0.77	0.44	0.56	0.73	0.45	0.56	0.81	0.51	0.63	0.51	0.29	0.37

Table A8: The average and standard deviation of precision, recall, and F-score of value predictions.

Method	Achievement			Power: resources			Security: personal			Security: societal		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
IS-each (BERT)	0.40±0.01	0.56±0.05	0.46±0.01	0.41±0.03	0.50±0.04	0.45±0.03	0.35±0.01	0.44±0.03	0.39±0.02	0.36±0.01	0.50±0.07	0.42±0.03
IS-shared (BERT)	0.37±0.02	0.29±0.05	0.32±0.04	0.38±0.02	0.26±0.06	0.31±0.04	0.32±0.08	0.20±0.08	0.24±0.08	0.41±0.03	0.34±0.05	0.37±0.04
IS-single (BERT)	0.38±0.05	0.39±0.05	0.38±0.04	0.40±0.01	0.33±0.03	0.36±0.02	0.37±0.02	0.40±0.06	0.38±0.03	0.38±0.03	0.36±0.06	0.37±0.02
IS-single (Llama)	0.41	0.39	0.40	0.44	0.29	0.35	0.46	0.41	0.33	0.45	0.31	0.37
DS-simple (BERT)	0.81±0.04	0.57±0.06	0.67±0.03	0.82±0.05	0.61±0.07	0.69±0.02	0.83±0.03	0.66±0.05	0.73±0.03	0.89±0.05	0.58±0.08	0.70±0.04
DS-sup (BERT)	0.78±0.03	0.62±0.04	0.69±0.01	0.84±0.03	0.58±0.04	0.69±0.02	0.83±0.04	0.64±0.06	0.72±0.03	0.84±0.04	0.65±0.06	0.73±0.02
DS-unsup (BERT)	0.65±0.05	0.60 ±0.03	0.62±0.03	0.73±0.05	0.55±0.04	0.63 ±0.02	0.65 ±0.06	0.72±0.05	0.68±0.04	0.71±0.06	0.67±0.06	0.68±0.02
DS-simple (Llama)	0.71	0.76	0.74	0.94	0.65	0.77	0.73	0.80	0.76	0.82	0.74	0.78
Baseline (random)	0.55±0.01	0.49 ±0.02	0.51±0.01	0.53±0.01	0.50 ±0.01	0.52±0.01	0.51 ±0.01	0.48±0.01	0.49±0.01	0.55±0.01	0.49±0.03	0.52±0.01

	Conformity: rules			Benevolence: caring			Benevolence: dependability			Universalism: concern		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
IS-each (BERT)	0.30±0.03	0.35±0.07	0.33±0.04	0.33±0.01	0.42±0.04	0.37±0.01	0.20±0.03	0.24±0.00	0.22±0.02	0.36±0.06	0.2±0.06	0.24±0.04
IS-shared (BERT)	0.28±0.03	0.18±0.06	0.22±0.04	0.28±0.02	0.18±0.03	0.22±0.03	0.14±0.03	0.15±0.06	0.14±0.04	0.32±0.04	0.16±0.10	0.20±0.08
IS-single (BERT)	0.30±0.05	0.26±0.04	0.27±0.04	0.35±0.03	0.27±0.06	0.30±0.04	0.23±0.02	0.19±0.02	0.21±0.02	0.37±0.02	0.20±0.05	0.25±0.04
IS-single (Llama)	0.42	0.19	0.26	0.35	0.22	0.27	0.33	0.13	0.18	0.35	0.17	0.23
DS-simple (BERT)	0.89±0.04	0.60±0.06	0.71±0.04	0.86±0.06	0.62±0.09	0.71±0.03	0.90±0.03	0.60±0.09	0.72±0.05	0.85±0.02	0.59±0.04	0.70±0.02
DS-sup (BERT)	0.89±0.01	0.61±0.03	0.72±0.02	0.89±0.04	0.56±0.08	0.69±0.04	0.86±0.03	0.67±0.06	0.75±0.03	0.85±0.04	0.59±0.07	0.69±0.04
DS-unsup (BERT)	0.70±0.07	0.68 ±0.06	0.69±0.04	0.78±0.05	0.52 ±0.05	0.62±0.04	0.66 ±0.05	0.55±0.06	0.60±0.03	0.70±0.06	0.59±0.06	0.64±0.03
DS-simple (Llama)	0.88	0.75	0.81	0.88	0.78	0.83	0.88	0.86	0.87	0.91	0.71	0.80
Baseline (random)	0.50±0.02	0.50 ±0.02	0.50±0.02	0.52±0.02	0.50 ±0.02	0.51±0.02	0.52 ±0.01	0.52±0.02	0.52±0.01	0.53±0.02	0.50±0.02	0.52±0.01

Table A9: Precision (P), recall (R), and F_1-score of the subjectivity prediction per value. STD shows the 5 runs on different train and validation sets with a fixed test set.