Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values

Neele Falk[1], **Gabriella Lapesa**[2,3]
[1] University of Stuttgart, [2] GESIS - Leibniz Institute for the Social Sciences, Cologne
[3] Heinrich-Heine University Düsseldorf
[1]neele.falk@ms.uni-stuttgart.de, [2]gabriella.lapesa@gesis.org

Abstract

The NLP community has converged on considering disagreement in annotation (or human label variation, HLV) as a constitutive feature of subjective tasks. This paper makes a further step by investigating the relationship between HLV and model uncertainty, and the impact of linguistic features of the items on both. We focus on the identification of moral foundations (e.g., care, fairness, loyalty) and human values (e.g., be polite, be honest) in text. We select three standard datasets and proceed into two steps. First, we focus on HLV and analyze the linguistic features (complexity, polarity, pragmatic phenomena, lexical choices) that correlate with HLV. Next, we proceed to uncertainty and its relationship to HLV. We experiment with RoBERTa and Flan-T5 in a number of training setups and evaluation metrics that test the calibration of uncertainty to HLV and its relationship to performance beyond majority vote; next, we analyze the impact of linguistic features on uncertainty. We find that RoBERTa with soft loss is better calibrated to HLV, and we find alignment between calibrated models and humans in the features (textual complexity and polarity) triggering variation.

1 Introduction

Since the perspectivist turn in NLP (Cabitza et al., 2023), disagreement in annotation, referred to as **human label variation** (Plank, 2022) in this paper, is no longer viewed as noise to be eliminated, but as a valuable source of information that should be expected and embraced. Human label variation (HLV) can be attributed to various factors such as the fuzziness and implicit nature of the target phenomenon, linguistic ambiguity, and the diverse socio-demographic backgrounds and lived experiences of the annotators.

This paper investigates the relationship between HLV and **model uncertainty**, i.e., the variation in model predictions as determined by inherent properties of the data. Building on Baan et al. (2024), we view model uncertainty as an indicator of variation in human perspectives that can be partially explained by the linguistic properties of a text. These properties may contribute to ambiguity, difficulty in interpretation, or noisy labels during training. Ideally, one would want to calibrate model uncertainty to human label variation, i.e., to align the probability distribution of the model with the diversity of human perspectives, an essential step towards developing fairer models.

Our investigation of the patterns of variation in human labels and model uncertainty focuses on the (automatic and human) annotation of two phenomena from social psychology: moral foundations (e.g. *care, fairness, loyalty*) (Graham et al., 2009, 2013) and human values (e.g. *be polite, be honest, have harmony with nature*) (Schwartz and Bilsky, 1987). Morals and values have garnered increasing interest in recent years, leading to the collection of annotated datasets, the application of LLM methods, as well as ethical reflection on the challenges related to the proper treatment of such culture-specific constructs (Vida et al., 2023). The interpretation of values and morality in text is highly subjective due to the lack of context and the fact that, due to the abstract nature of these concepts, annotators rely on their own morals and values to interpret the text.

A crucial component of our approach is the use of **linguistic features** as an interpretation tool for both HLV and model predictions. Consider the tweet *"Every man needs to remind @realDonaldTrump what respect and integrity mean. Humility and compassion are necessary to serve as a leader."*, annotated with *care* by two and *fairness* by one annotator(s) (Fig. 1). Annotation might be influenced by how annotators perceive named entities like "Donald Trump" and their prior knowledge about him, as well as their comprehension of abstract concepts like empathy and humility

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22898–22921

and the different associations individual annotators have with these concepts. In addition, the political ideology and moral or value preferences of the annotators complement the variation that can be explained by textual properties. Given that informations about annotators are missing in the current datasets, we focus on the linguistic properties which can, in a certain sense, serve as an approximation of it (i.e., knowledge of named entities approximates world knowledge).

In the first part of the paper, we start by analyzing the observed HLV in three established datasets. We enrich them with features related to linguistic properties of the items: complexity (i.e., readability, length, abstractness), polarity (i.e., sentiment, emotion), lexical choices (i.e., the use of named entities or of moral-specific lexica), and pragmatic phenomena (i.e., irony, sarcasm, storytelling). Next, we employ regression analysis to mine HLV based on linguistic properties, thereby addressing our first research question, **RQ1: what drives HLV in annotating morals and values from text?**

In the second part of the paper, we evaluate how well the probability distribution of classification models aligns with the distribution of human judgments and to what extent the predictions match the majority label or individual annotators' decisions. We experiment with RoBERTa and flan-T5, trained on the majority vote, compared to approaches that calibrate the respective model to HLV. We evaluate both calibration and classification performance using a range of metrics that extend beyond majority-based evaluation, with a focus on aleatoric uncertainty arising from supervision, thus addressing our second research question: **RQ2: how can we calibrate model uncertainty to HLV?** Finally, we investigate the impact of linguistic factors on uncertainty and answer our last questions: **RQ3a: What linguistic factors influence model uncertainty? RQ3b: Are models and humans influenced by similar patterns?**

We find that RoBERTa can be calibrated well with soft loss and outperforms LLMs and majority vote models (RQ2). Complexity and polarity significantly impact both HLV (RQ1) and model uncertainty (RQ3a); additionally, calibrated models exhibit similar patterns to HLV (i.e., they are influenced by the same features) compared to the non-calibrated ones (RQ3b).

The contributions of our work are manifold. At the level of the **research on morals and values**, we fill two crucial research gaps: a) while HLV is a constitutive feature of morals and values, research on it has been limited so far; b) We are the first to integrate calibration to HLV in the computational modeling of these phenomena. At the level of **analysis**, we contribute to a better understanding of the linguistic properties and related annotation patterns in three reference datasets. At the **modeling level**, the contribution of our experimental setup (two tasks, three datasets, nine evaluation metrics, and two model architectures) goes beyond the specific phenomena we focus on. Calibration in LLMs is understudied: our findings can guide future work in this domain, and towards the development of more interpretable, trustworthy and fair models.[1]

2 Related Work

Human Label Variation and Model Uncertainty Recent progress in NLP resulted in a shift towards embracing disagreement and subjectivity, rather than reducing it (Aroyo and Welty, 2015; Cabitza et al., 2023; Fleisig et al., 2024). In this context, modeling approaches are developed that incorporate human disagreement as a valuable source of information, e.g., predicting disagreement as an auxiliary task (Romberg, 2022), applying soft loss learning methods (Uma et al., 2021; Leonardelli et al., 2023) which learn to match the underlying human distribution directly, or focussing on annotator-based modeling (Mostafazadeh Davani et al., 2022; Gordon et al., 2022). Several works investigate the factors which influence human disagreements: linguistic factors (Pavlick and Kwiatkowski, 2019), socio-demographics (Wan et al., 2023; Prabhakaran et al., 2024), or attitudinal information (Jiang et al., 2024). Other works focus on the impact of disagreement on model performance (Leonardelli et al., 2021). However, the impact of incorporating this information on model performance depends on the data, the task, which information helps to explain annotation variance (Beck et al., 2024; Hu and Collier, 2024), and how 'performance' is measured.

In addition to quantifying model performance through traditional metrics (which is still an open question in subjective tasks that deal with HLV), another aspect of performance is calibration. In the context of HLV, one perspective is to measure whether the models probabilities align with human label variation (Baan et al., 2022). This perspective can complement the notion of calibration when

[1] Code, datasets with features, and regression outputs are available on https://shorturl.at/ykiym

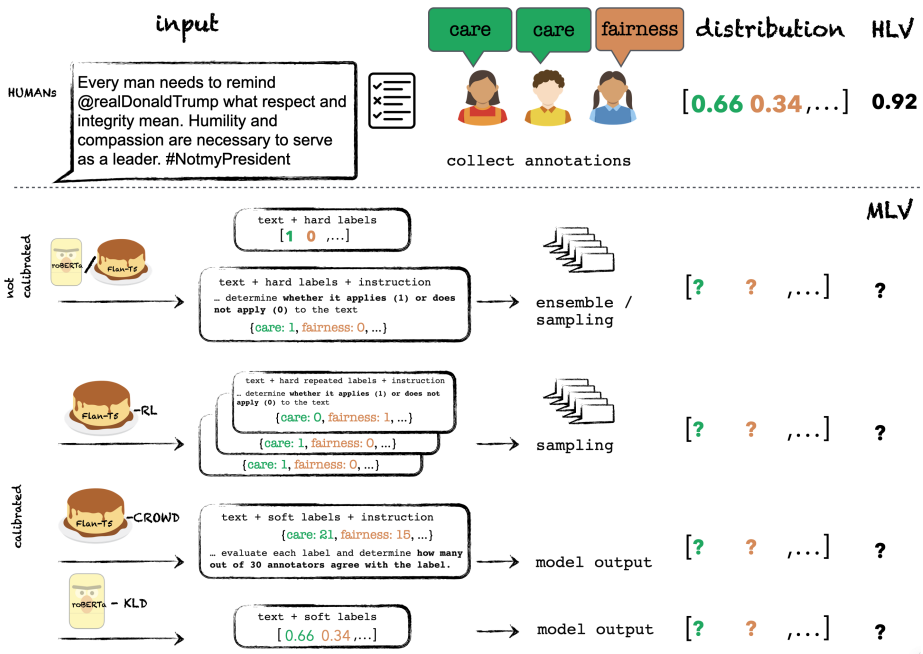

Figure 1: **Upper Panel**: Example tweet from the Moral Foundations Twitter Corpus (MFTC) with annotator disagreement. The distribution shows the proportion of annotators selecting each foundation. HLV (Human Label Variation) is the normalized entropy averaged over labels. **Lower Panel**: Models are either calibrated (trained with soft or repeated labels) or uncalibrated (trained on majority vote). After training, soft labels are produced directly, via sampling, or using ensembles. MLV (Model Label Variation) is the normalized entropy across labels. We assess model-human alignment by measuring distance between distributions or correlation between HLV and MLV.

looking at model confidence: does the model's accuracy correspond with its uncertainty? According to Baan et al. (2024), both perspectives on model uncertainty are equally important for fair and reliable NLP systems. This work focusses on model uncertainty from the perspective of HLV and investigates the relationship between model's probability distribution and the human distribution over labels.

Modeling morals and values The modeling of morals and values has been explored in various domains, e.g., political discourse (Johnson and Goldwasser, 2018; Islam and Goldwasser, 2022), argumentation (Alshomary et al., 2022), stories (Wu et al., 2023; Hobson et al., 2024). Modeling approaches range from dictionary-based (Graham et al., 2009; Hopp et al., 2020) to supervised classification (Lin et al., 2018; Kobbe et al., 2020; Hoover et al., 2020), to LLMs (Roy et al., 2022; Asprino et al., 2022). Frequently encountered problems are the implicitness and subjectivity of these phenomena and, at the modeling level, class imbalance, out-domain generalization (Liscio et al., 2022) and interpretability (Liscio et al., 2023). For a survey of NLP-based approaches to automatically identify morals in text, refer to Reinig et al. (2024).

Fewer works have looked into HLV in annotating morals and values, while both have been considered as a source for HLV in other target phenomena, e.g. annotator's morals for variance in offensiveness detection (Mostafazadeh Davani et al., 2024), or user's human values for disagreements in online discussions (van der Meer et al., 2023). Van Der Meer et al. (2024) investigate how a model can be calibrated to match HLV in an active-learning scenario when classifying moral foundations in the moral foundation Twitter corpus (MFTC) (Hoover et al., 2020). Beyond majority-vote performance, they introduce two F1 metrics that equally consider all annotators or measure F1 on minority perspectives, which we incorporate into our evaluation. Mokhberian et al. (2022) has looked into disagreement in moral foundations and evaluated the impact of removing it from training data, considering it as noise. The work that is most related to ours is by Alvarez Nogales and Araque (2024), who investigated HLV in the MFTC and developed annotator-based models trying to explore the performance on individual annotators in different domains; they also developed a classifier to identify textual instances that are hard to annotate (result in high disagreement) and employed an interpretability method to explore the lexical features picked up by their model.

3 Datasets

We conduct our experiments on three standard datasets for moral foundations and human values.[2]

MFTC The Moral Foundation Twitter Corpus (Hoover et al., 2020): 35k tweets, annotated according to the five-factor taxonomy of the Moral Foundation Theory (MFQ5: *care, fairness, purity, authority, loyalty*), by 23 trained annotators.

MFRC The Moral Foundation Reddit Corpus (Trager et al., 2022): 16k Reddit posts, annotated by 5 trained annotators according to the six foundations in the updated Moral Foundation Theory (MFQ6) (Atari et al., 2023). With respect to the MFQ5 version, *fairness* is split into two foundations: *equality* and *proportionality*. Additionally, a *thin morality* label is introduced for items that exhibit a weak morality signal, not categorizable in the other labels but also to be distinguished by the "non moral" label.

Human Values (HV) The Touché23-ValueEval dataset: 9k arguments from diverse domains, annotated for human values by 38 annotators, according to the multilayer taxonomy by Kiesel et al. (2022).

While detailed guidelines were used for MFTC and MFRC, annotators were also encouraged to follow their own intuition, leaving room for subjectivity, the guidelines for HV are more prescriptive.

4 Human Label Variation

4.1 Methodology

4.1.1 Quantifying Disagreement

For each instance in each dataset, we quantify the amount of HLV as the normalized entropy on all labels (moral foundations or human values). We have 6 moral foundations for MFTC (5 foundations + non-moral), 8 for the MFRC (6 foundations + non-moral + thin morality) and 20 for human values (we use the 20 coarse-grained values that have also been used in the Shared Task). For each class, we calculate the soft label as the relative amount of annotators that selected it. Since each instance is annotated in a multi-label fashion (one or several values or foundations can be present) we calculate the normalized entropy of each dimension first, and take the average to retrieve a mean entropy score for the instance. Appendix Fig. 6 visualizes the relative amount of data expressing different degrees of variation (from low to very high entropy). We can see that MFTC displays the highest amount of variation, with more than half of the data showing a medium to high disagreement. For all datasets, the majority of instances express at least some (weak) disagreement.

4.1.2 Linguistic Features

We hypothesize that some of the HLV can be explained by textual properties of the instances. These properties make interpretation harder (complexity, vagueness, ambiguity) or tap into the subjective nature of the task (emotions, sarcasm, entities). We consider features from four groups.[3]

Linguistic complexity: these features express the diversity in vocabulary use, the readability, concreteness or specificity or more surface-related properties that make a text more complex (type-token ratio, amount of long words). We expect disagreement to be higher when the text is more complex and more abstract.

Polarity: features related to emotions (surprise, anger) or sentiment and polarity (valence, dominance, arousal.) These features were extracted with a lexicon-based approach. We expect texts that express a stronger polarity to trigger higher disagreement because of the additional level of subjectivity that may be tricky for the annotators to disentangle from the content-level, which is the target of the annotation. Different annotators may be subject to the interference of polarity to a different extent.

Lexical: we investigate the amount of loaded words (i.e., words that are strongly associated with certain values or moral foundations), as well as the amount of named entities. We expect those words to trigger certain annotators to assign a certain dimension, which could either lead to higher agreement (in case those words trigger annotators equally) or higher disagreement (e.g., in cases where moral sentiment is associated with a certain named entity because of world knowledge – this could affect some annotators but not others).

Pragmatic: we investigate features of pragmatic phenomena: sarcasm, irony, and whether the text contains a story or personal experience. These features are extracted with off-the-shelf classifiers. We expect that sarcasm/irony may make the true reading of a text more difficult to interpret, leading to higher disagreements. Additionally, morals and values can be implicitly expressed through stories and personal experiences, again, creating an addi-

[2] For more details on the datasets, including preprocessing steps, refer to Appendix A.

[3] Details on feature extraction are provided in Appendix B

	Human Values	MFRC	MFTC
Complexity	1.984	4.041	10.463
Polarity	0.857	3.755	3.855
Pragmatic	0.038	0.356	0.930
Lexical	—	1.583	2.705
Total	2.879	9.735	17.953

Table 1: Regression analysis of disagreement. For each feature category, we report the sum of explained variance for features that belong to that category. We also report the total amount of explained variance stemming from linguistic factors for each of the three datasets.

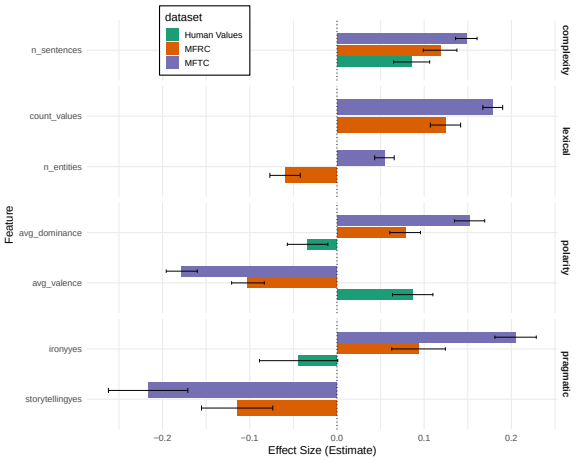

Figure 2: Estimated effect sizes and directions for the most explanatory features, selected via AIC and grouped by feature type. Each color represents a separate regression model (one per dataset). If a color is missing for a feature, it indicates that the feature did not have a statistically significant effect in the corresponding model.

tional interpretation step which may be handled differently from different annotators.

4.2 Analysis: Linguistic Features and HLV

To mine the impact of linguistic features on HLV, we train a regression model predicting the average entropy of an instance given the linguistic features described in section 4.1.2. For feature selection and interpretation of results, we follow the procedure detailed in Appendix C.

Table 1 displays the result of each regression model (one per dataset) with the total amount of explained variance and the amount of variance that can be ascribed to the respective feature type. Comparing the total amount of variance that can be explained by the linguistic features we notice a stark difference between the datasets: little variance can be explained for the Human Values dataset (∼3%), more for the MFRC (∼9%) and most for MFTC (∼18%). Given that the MFRC and the MFTC follow similar annotation guidelines and handle subjectivity in a similar way, we hypothesize the difference between these two for the following reasons: (a) tweets are shorter and more ubiquitous, so harder to annotate in general, while Reddit posts provide more context and not as many hashtags / references that require background knowledge; (b) there is a smaller annotator pool for MFRC; and (c) the annotation of MFRC allows very weak or vague signals of morality to be labeled as *thin Morality* opposed to having to assign it a specific foundation. Regarding the lower variance within the HV corpus we hypothesize that the items are more well-formed and less likely to express subjective and vague language (they were pre-selected based on certain quality criteria) and the guidelines were more prescriptive than for the other two datasets.

We now turn to the inspection of the estimates of the regression model, displayed in Figure 2. Recall that since we are predicting entropy, a positive estimate (higher entropy for higher values of the feature of interest) indicates higher disagreement.

As shown in Table 1, complexity is the strongest predictor of variation. The direction of the effects confirms the hypothesis: **more complex texts**, as measured, for instance, by the number of sentences (`n_sentences`), trigger more variation.

Turning to the polarity features, we also find expected **positive effects for the presence of emotions or stronger sentiment**. Interestingly, we find opposite effects when comparing valence and dominance between datasets: high valence increases disagreement in human values but reduces it in moral foundations, while high dominance does the reverse. It is possible that certain polarity features are strongly associated with specific foundations or values (e.g., anger with unfairness, joy with achievement). If a feature strongly associated to a foundations, it should increase agreement; if it is mere subjectivity, it should increases noise. A future step could be to conduct the analysis per foundation or value.

As for the lexical features, they have a significant impact only on the moral foundation datasets. **Named entities** (`n_entities`) behave differently in the two corpora: they **trigger disagreement in the Reddit** corpus (MFRC), **support agreement in the twitter corpus** (MFTC). Given the brevity of tweets, named entities often dominate the message, indicating an actor or event location that makes the moral easy to interpret. In contrast, Reddit texts,

with their different lengths and purpose, show a less clear relationship between entities and morals.

As for pragmatic features, while **irony predicts higher disagreement** for the moral foundation corpora, **storytelling decreases disagreement**, suggesting that morals are explicitly encoded in the stories, and thus easier for annotators to agree upon.

5 Model Uncertainty

5.1 Methodology

5.1.1 Models

We use two different architectures: RoBERTa (Liu et al., 2019) for traditional text classification and flan-t5 (Chung et al., 2024) as an LLM-based approach. For each architecture, we compare a model using (majority) hard labels with variants that incorporate either the human distribution or the individual annotator perspectives. In what follows we list the models we employ and how we retrieve model uncertainty and a probability distribution, Fig. 1 depicts an overview of the models:[4]

RoBERTa-mv: we fine-tune RoBERTa-base using the **majority vote** for each instance as a gold label and the binary-cross entropy loss in a multi-label fashion. To receive a distribution of probabilities for each label, we employ a deep ensemble (Lakshminarayanan et al., 2017): we train five model instances with five different random initializations and extract the probability distributions by taking the mean over predicted probabilities.

RoBERTa-KLd: we train a RoBERTa-base model with a soft loss. The target is to reduce the Kullback-Leibler divergence (KLd) between the probability distribution predicted by the model and the distribution observed from human data, specifically for each label. The total loss is then defined as the average divergence over all the labels.

flan-t5-mv: we write a detailed instruction for the model and ask it to generate a binary vector, where each element represents the presence or absence (1 or 0) of a corresponding label. We provide the model with a mapping of index to label and task the model to strictly keep the correct label ordering. To retrieve a probability distribution for each label, we sample 30 generations during inference (temperature = 0.7). The relative frequencies gathered for each class are used to estimate the distributions.

flan-t5-RL: one of the simplest approaches of incorporating the information about disagreement into the training data is **repeated labeling**: we use all input-annotation pairs that are available for a specific item in training. This is feasible since all our datasets consist of items with 3-5 annotations. This approach is an intuitive way to inject multiple annotator-perspectives into the LLM. Similarly to flan-t5-mv, we use sampling to retrieve a probability distribution for each label. We expect that this model will produce output aligned with the annotators encountered during training, thus serving as a better proxy for the human distribution.

flan-t5-crowd: Instead of fine-tuning an LLM to generate precise probability distributions like RoBERTa-KLd, which would be sub-optimal, since the model would need to generate fine-grained floating point numbers, we instruct the model to predict the number of annotators agreeing with a label. As datasets vary in annotator count, the model estimates this number out of 30 (corresponds to the sampling size for the LLM-based approaches). We map the original soft label from each dataset to a number within this range (0 to 30) (e.g., if 2 out of 4 annotators agreed on the label, we set it to 15).

5.1.2 Evaluation metrics

Calibration to HLV With the following four metrics we want to answer the question: **How well does the model distribution align with the human distribution?** The human gold distribution is retrieved based on the relative frequency of the positive / negative label for each class.

JS-divergence measures the distance between a gold and predicted distribution, serving as a symmetric, smoothed version of KL-divergence. We calculate this distance per label and instance, then average across all labels and instances to obtain an aggregated JS-score for a model on a dataset.

TVD: the total variation distance (Baan et al., 2022) measures the largest possible difference in the probabilities that the two distributions assign to the same events. It is more intuitive than the JS-divergence and less sensitive to small differences. We retrieve the instance-based TVD for each label and then report the mean over the labels and all instances in a particular dataset.

EntCE: the human entropy calibration error (Baan et al., 2022) measures the absolute difference between the entropy scores of humans and models. This score tells us whether a model is overconfident (> 0), aligned ($= 0$) or uncertain when humans agree (< 0). Similarly to JSD and TVD, this is an instance-based metric. The overall entCE is the

[4] Implementation and training details are in Appendix E.

mean of all the labels in all instances.

Corr(ent) calculates the spearman correlation between the human and model entropy, thus representing an overall alignment metric, compared to the other instance-based metrics.

Performance and alignment with annotators With our F1-based performance metrics we would like to quantify: **how well do the predicted labels align with the majority vote, individual annotators, and the least aligned annotators?** More specifically, we consider the following metrics.

F1-macro: we look at the F1-macro score to get an idea of how well the predicted label of a model aligned with the majority vote.

Annotator-based: For better representing different annotators we want to look at the performance compared to individual annotators. We evaluate model performance using each annotator's labels as ground truth (*F1-per-annot*) and consider the average F1-scores of annotators within the lowest 20th percentile (*worst-off f1*, Van Der Meer et al. (2024)). This is especially informative when dealing large pools of annotators, where individual performance significantly differs from the majority.

For all metrics, we calculate significance with the Almost Stochastic Order test (del Barrio et al., 2018) as implemented by Ulmer et al. (2022).

5.2 Modeling results

Calibration to HLV Comparing all models on all datasets (Table 2) we find that the distance-based approaches (RoBERTa-KLd and flan-t5-crowd) often achieve lower distances to the target distribution, better entropy calibration, and higher entropy correlation. This suggests that explicitly training models to approximate the target distribution effectively aligns model and human probabilities. However, retrieving uncertainty from an ensemble (RoBERTa-mv) or with sampling (flan-t5-mv) is a strong baseline, in some cases well aligned with human probabilities or entropy (e.g., on MFRC).

How do models differ in alignment when we compare low-disagreement vs. high-disagreement instances? Given that some of the metrics reflect alignment between models and humans at the instance level, we want to compare their overall distribution to better understand model behavior related to uncertainty calibration against HLV. We plot the distribution of the TVD across the different instances for two subsets of the data: instances with low variation and high variation (high entropy). Figure 3 shows the TVD scores for MFRC (Appendix Figure 10 covers all datasets), contrasting the non-calibrated model with its calibrated counterpart for flan-t5 and RoBERTa.

Setup	JSD	TVD	entCE	corr(ent)
DATASET: HV				
r-mv	0.165 ± 0.004	0.134 ± 0.002	-0.144 ± 0.019	0.155 ± 0.012
r-KLd	**0.142 ± 0.003**†	**0.122 ± 0.002***	**-0.068 ± 0.010**	**0.253 ± 0.022***
f-t5-mv	0.166 ± 0.006	0.139 ± 0.004	-0.112 ± 0.004	0.070 ± 0.034
f-t5-RL	0.152 ± 0.016	0.132 ± 0.001	-0.107 ± 0.031	**0.144 ± 0.087**†
f-t5-cr	**0.138 ± 0.002***	0.153 ± 0.002	0.093 ± 0.005	0.092 ± 0.024
DATASET: MFRC				
r-mv	0.122 ± 0.001	0.113 ± 0.001	0.109 ± 0.010	0.291 ± 0.023
r-KLd	0.124 ± 0.015	**0.103 ± 0.005***	-0.087 ± 0.055	**0.407 ± 0.024***
f-t5-mv	**0.114 ± 0.004**†	**0.110 ± 0.002**†	0.049 ± 0.022	**0.351 ± 0.008**†
f-t5-RL	0.175 ± 0.008	0.144 ± 0.006	-0.189 ± 0.024	0.074 ± 0.113
f-t5-cr	0.117 ± 0.017	0.120 ± 0.019	**0.002 ± 0.006**	0.202 ± 0.087
DATASET: MFTC				
r-mv	**0.139 ± 0.002**†	0.119 ± 0.003	0.076 ± 0.050	**0.505 ± 0.040**†
r-KLd	0.153 ± 0.002	**0.117 ± 0.002**†	-0.099 ± 0.013	**0.535 ± 0.018***
f-t5-mv	0.165 ± 0.057	0.154 ± 0.049	-0.05 ± 0.187	0.322 ± 0.209
f-t5-RL	0.238 ± 0.004	0.219 ± 0.003	-0.307 ± 0.015	0.003 ± 0.014
f-t5-cr	**0.116 ± 0.005***	0.118 ± 0.006	**0.025 ± 0.004**	0.478 ± 0.023

Table 2: Calibration Results for Human Values (HV), MFRC and MFTC. Abbreviations: r=RoBERTa; f=flan; f-t5-cr=f-t5-crowd. Best results are bolded. * denotes that the model is stochastically dominant over all other models. † indicates the model is stochastically dominant over three other models ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.5$)

All datasets exhibit similar trends in the high entropy scenario: soft approaches (calibrated models) lead to a more concentrated distribution at lower TVD scores with lower quartiles and a tail towards the lower end. This indicates improved uncertainty calibration, with fewer high-TVD instances compared to non-calibrated models. The datasets differ in how the approaches succeed in the low entropy scenario (when annotators agreed). For HV, the models perform similarly, with RoBERTa-KLd aligning slightly better with high-agreement instances. For MFTC, the calibrated flan-t5 outperforms the non-calibrated flan-t5, but this is not the case for RoBERTa. Here, we can also observe many outliers with high TVD scores. As shown by Figure 3, in the MFRC the soft approaches are not well calibrated with high agreement instances: models maintain higher uncertainty, even when humans agree, shifting from overconfidence (usually the problem in non-calibrated models) to excessive uncertainty. This highlights the need for stronger regularization during training, either through early stopping or incorporating a loss function that differ-

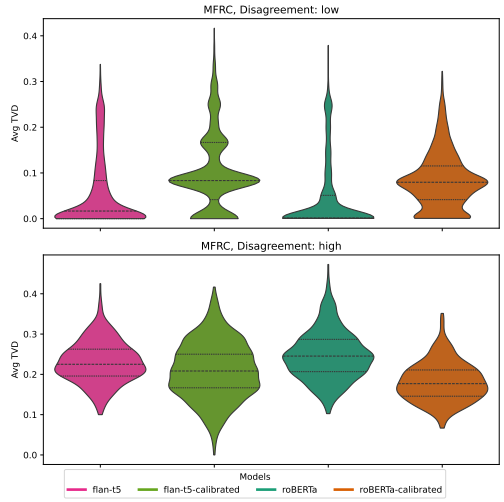

Figure 3: Violin plots for the TVD scores comparing low-disagreement and high disagreement scenarios. Comparing models without and with calibration for the MFRC. Lower TVD values indicate better alignment.

Setup	f1-macro	f1-per-annot	worst-off-f1
DATASET: HV			
r-mv	**0.639 ± 0.002***	**0.619 ± 0.002***	**0.519 ± 0.005†**
r-KLd	**0.634 ± 0.005†**	**0.615 ± 0.006†**	**0.526 ± 0.006***
f-mv	0.492 ± 0.017	0.494 ± 0.017	0.437 ± 0.014
f-RL	0.512 ± 0.035	0.511 ± 0.032	0.451 ± 0.025
f-dist	0.504 ± 0.004	0.500 ± 0.002	0.460 ± 0.005
h-majority	–	0.795 ± 0.00	0.632 ± 0.01
DATASET: MFRC			
r-mv	**0.730 ± 0.009***	**0.727 ± 0.011***	**0.665 ± 0.013***
r-KLd	**0.712 ± 0.013†**	**0.710 ± 0.014†**	**0.620 ± 0.021†**
f-mv	0.643 ± 0.017	0.639 ± 0.014	0.525 ± 0.023
f-RL	0.495 ± 0.023	0.494 ± 0.023	0.458 ± 0.016
f-dist	0.612 ± 0.050	0.611 ± 0.05	0.545 ± 0.032
h-majority	–	0.745 ± 0.00	0.648 ± 0.01
DATASET: MFTC			
r-mv	**0.828 ± 0.003***	**0.733 ± 0.007***	**0.540 ± 0.034†**
r-KLd	**0.821 ± 0.002†**	**0.728 ± 0.004†**	**0.547 ± 0.037***
f-mv	0.681 ± 0.181	0.618 ± 0.135	0.484 ± 0.057
f-RL	0.463 ± 0.006	0.455 ± 0.008	0.402 ± 0.021
f-dist	0.788 ± 0.014	0.698 ± 0.015	0.544 ± 0.027
h-majority	–	0.650 ± 0.00	0.531 ± 0.01

Table 3: Performance Results for Human Values (HV), MFRC and MFTC. Abbreviations: r=RoBERTa; f=flan; h-majority=human majority. Best results are bolded. * denotes that the model is stochastically dominant over all other models. † indicates the model is stochastically dominant over 3 other models ($\epsilon_{\min} < \tau$ with $\tau = 0.5$)

entiates between high- and low-entropy cases (e.g., the approach by Baumler et al. (2023)).

Performance and alignment with annotators Table 3 displays classification performance, as well as the performance between the human majority vote label and each annotator and the respective average over the 20 percent worst-off annotators (*human majority*). As in previous work (Alvarez Nogales and Araque, 2024), we observe that the majority vote does not represent all annotators equally well. The F1-scores range between 0.63 and 0.99 (HV), 0.62 and 0.83 (MFRC) and 0.49 and 0.75 (MFTC). These discrepancies, up to 30% among annotators, suggest that standard evaluation falls short in equally representing all annotators.

Comparing the different models, we find that RoBERTa outperforms the instruction-fine-tuned approaches on all datasets. Between soft and majority vote (mv) approaches we find a mixed picture: For RoBERTa the mv-based approach represents the majority well, in some cases the soft approach outperforms the ensemble for representing the 20 percent worst-off annotators, thus being more fair. Repeated labeling (RL) significantly diminishes performance for flan-t5 across datasets. Combined with its poor calibration, this method is inefficient, inadequately calibrated, and yields subpar results for disagreement-aware modeling of morality. An exception is the HV dataset, where flan-t5-RL marginally surpasses flan-t5-mv in both calibration and performance. Conversely, flan-t5-crowd, the other soft approach, achieves significantly better results on certain metrics, particularly the per-annotator-based ones (worst-off across all datasets, per-annot on 2 out of 3 datasets). Therefore, we find this soft method for flan-t5 to be better calibrated, more effective, and more fairly reflecting annotators compared to the mv-based approach.

Summary Considering all calibration and performance metrics, RoBERTa-KLd is the best model (and most robust across all datasets), with the additional benefit of being also the most efficient (compared to ensemble and LLM-based approaches).

5.3 Analysis: Linguistic features and Model Uncertainty

We now analyse model uncertainty with respect to linguistic features. We want to to a) identify the feature types that impact model uncertainty the most and b) know whether good calibration to HLV also corresponds to the same features impacting model uncertainty. We apply the same method as in Section 4.1.2, predicting model uncertainty which we define as the normalized entropy we retrieve from the models' predicted probabilities. Higher entropy corresponds to higher uncertainty.

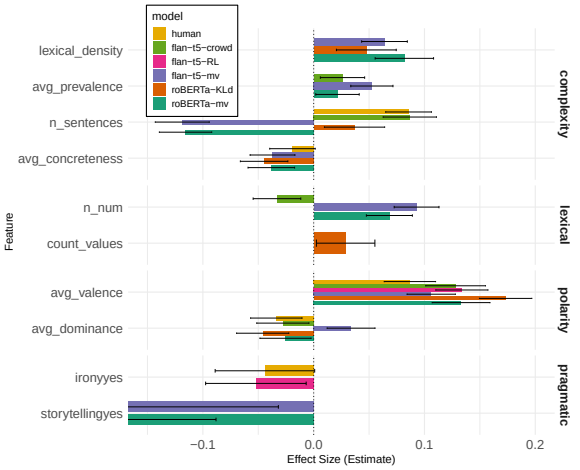

Figure 4: Human values – regr. analysis of model uncertainty: estimates and direction of effects, most explanatory models selected by AIC, grouped by feature type

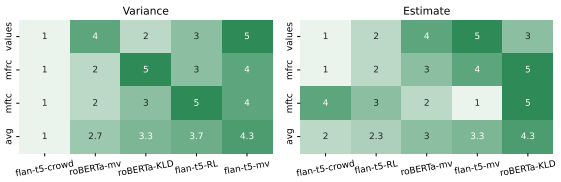

Figure 5: Ranking of models when comparing influential features on model uncertainty with features impacting HLV. Lower ranks = more alignment.

What linguistic factors influence model uncertainty? Appendix tables 6, 7, and 8 show the explained variance for each model type for each dataset. For MFRC and HV, models show higher explained variances, meaning linguistic features affect model uncertainty more than they do for humans (e.g., for HV 5–16% for models compared to roughly 3% for humans). The RL method shows the least amount of explained variance, indicating less influence from linguistic factors. Even without calibration, uncertainty of majority-vote methods is impacted by linguistic features, with the MFRC corpus showing a peak of 18% for the flan-t5-mv model. For all models, textual complexity most significantly influences model uncertainty, followed by polarity features, both feature types being tied to vagueness and ambiguity. Conversely, pragmatic features are subtler and perhaps more pertinent to human subjectivity in interpreting narratives or irony, especially in moral contexts.

The effects displayed in Figure 4 for HV, and Appendix Figures 8 and 7 for MFRC and MFTC respectively, show similar directions for model and humans (with different effect sizes). Most exceptions occur in the HV dataset in which we find additional complexity features impacting uncertainty (e.g. lexical density and prevalence) or the presence of numbers (`n_num`) or loaded words (`count_values`). We also find differences between calibrated and non-calibrated models (opposite directions of effects): longer texts predict a lower model uncertainty for non-calibrated models, while the effect is opposite for calibrated models (in alignment with humans). Other differences are higher uncertainty for RoBERTa-Kld when the texts contain value-loaded words, lower uncertainty for the mv-methods when a story is present in the text, while RL is the only approach that is affected by irony (lower uncertainty when a text is ironic).

Are models and humans influenced by similar patterns? To assess the alignment between humans and models regarding the influence of textual features, we represent the results of each regression analysis as feature vectors, using either the **explained variance** of each feature or the **effect estimates**. For each dataset, we compute the Manhattan distance between the human vector and the corresponding model vectors. Figure 5 shows the ranked models based on similarity to humans in their feature patterns. Flan-T5 (LLM-based approach) exhibits feature patterns more similar to humans but only when explicitly calibrated (crowd).

6 Conclusion

This work investigated the alignment between humans and models in the annotation of morality and values: do humans and models align in their probability distributions and annotated labels in these tasks? We found that soft approaches are more aligned with humans and that a traditional RoBERTa classification model with a soft loss outperforms the other models with respect to both calibration to HLV and classification performance.

In our final analysis, we looked at the impact of linguistic features on model uncertainty in comparison to humans. In many cases models are more influenced by textual properties than humans: humans lived experiences and moral or value-based preferences should account for more variance in annotation, while models are more sensitive to textual cues. Still, both show similar patterns: higher complexity, polarity, loaded lexical cues, and complex pragmatic implications lead to greater uncertainty.

7 Limitations

This paper looks at linguistic factors that influence human disagreement and model uncertainty. What is still missing is a more complete picture of other factors that influence the disagreement (e.g. the annotators personal background and preferences regarding moral foundations and human values). On top of that, more analysis on the relationship between individual human uncertainty and uncertainty as disagreement should be conducted, which eventually also requires more data collection.

For future work, we recommend collecting new data with annotator background information and specific human confidence levels for each instance. This additional information may clarify how annotator background factors contribute to disagreement, while individual confidence scores can help to draw a more complete picture of human uncertainty.

Up to date the community lacks resources of texts annotated with morality and values in other languages, especially from non-WEIRD countries. This study is based on three large and established datasets, all of which are in English.

Regarding the modeling, this paper focuses on two model architectures and versions that are calibrated on human disagreement (soft approaches). These architectures are RoBERTa and flan-t5. It would be interesting to investigate more model architectures and see how their patterns in uncertainty differ from the ones we analyze here. In this line, there are also more approaches to calibrate models on human disagreement. While more approaches have been investigated in traditional transformer-based text classification, this research direction is still new and explicitly aligning LLMs and human probabilities during training has not yet been fully investigated. Different prompting strategies or a reinforcement learning stage would be possible directions to further explore methods for calibration.

Therefore, more research is needed (a) to develop a robust method to extract uncertainty from LLMs and (b) to calibrate them more to human label variation. Our results show that the more established method works better on this task, but more methods and LLMs need to be evaluated to maximize and fully evaluate their effectiveness in that direction.

8 Ethics

This work explores how textual features affect model uncertainty in order to enhance algorithmic transparency which is crucial for neural networks and LLMs as their applications in societal contexts increases. A major risk of inferring morality or human values from text lies their application in user profiling, e.g., for the purpose of personalized advertising. Models that are able to generate morally framed content can be applied for manipulation, e.g. in debates.

Another important ethical aspect of this work is the evaluation of models with respect to which perspectives are well represented. We find that models performance varies largely between annotators. Calibrating models on human disagreement makes them more fair (e.g. better F1 scores for all annotators). As suggested by the perspectivism paradigm, more research effort needs to be put into perspective-based modeling in NLP and evaluation methods that help to understand who (e.g. what annotator) is favored by certain model types in certain tasks.

Acknowledgments

We would like to thank Ana Barič and the anonymous reviewers whose feedback helped us improve this paper. This research has been partially funded by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation).

References

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.

Anny D. Alvarez Nogales and Oscar Araque. 2024. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 67–77, Torino, Italia. ELRA and ICCL.

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

Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovi. 2022. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 33–41, Dublin, Ireland and Online. Association for Computational Linguistics.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. 2018. *An Optimal Transportation Approach for Assessing Almost Stochastic Order*, page 33–44. Springer International Publishing.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

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

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*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. *Moral Foundations Theory*, page 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani,

Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2020. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Tunazzina Islam and Dan Goldwasser. 2022. Understanding COVID-19 vaccine campaign on facebook using minimal supervision. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 585–595. IEEE.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Jana Juroš, Laura Majer, and Jan Snajder. 2024. LLMs for targeted sentiment in news headlines: Exploring the descriptive-prescriptive dilemma. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 329–343, Bangkok, Thailand. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

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

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

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, page 552–559. IEEE.

Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn Jonker, Kyriaki Kalimeri, and Pradeep Kumar Murukannaiah. 2023. What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14113–14132, Toronto, Canada. Association for Computational Linguistics.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep 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.

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.

Maximilian Maurer. 2025. Elfen - efficient linguistic feature extraction for natural language datasets. https://github.com/mmmaurer/elfen.

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

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.

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

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. 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.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. 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 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A survey on modelling morality for text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand. Association for Computational Linguistics.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards few-shot identification of morality frames using in-context learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196, Abu Dhabi, UAE. Association for Computational Linguistics.

Shalom H. Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3):550–562.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus. *CoRR*, abs/2208.05545.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance - easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint*.

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.

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.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

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

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Appendix

A Datasets and pre-processing

MFTC The Moral Foundation Twitter Corpus (Hoover et al., 2020) consists of 35k tweets extracted from seven controversial discussion threads (e.g., Black Lives Matter, 2016 presidential election). The data is annotated according to the five-factor taxonomy of the Moral Foundation Theory (*care, fairness, purity, authority, loyalty*) and each factor can be labeled as a vice (something desirable) or a virtue (something to avoid), for example, a text might express the desire to protect someone (*care*) or preventing them from being harmed (*harm*). The 23 annotators, all undergraduate researchers, received training specifically for the task but were advised not to rely excessively on heuristics to achieve high agreement, but to still follow their intuition. The authors acknowledge that, despite efforts to minimize annotator bias, inferring moral values from text is inherently subjective. They chose not to artificially reduce this subjectivity by resolving disagreements.

For our experiments, we merge annotations for vice and virtue into one label to indicate the presence / absence of a particular moral foundation. We drop duplicates and only keep items that have at least three annotations. Most of the items have been annotated by three or four annotators. After pre-processing, 33,684 instances remain. For the majority vote, we set the foundation label to 1 if at least 50 percent of the annotators agreed with that label. If there is a tie between *non-moral* and a specific label, we set the majority vote to the foundation and *non-moral* to 0. We found this case in 1354 instances. If there is no majority vote (no 50 percent for any label and not for *non-morality*) we label the instance as *non-morality*. This was the case for 2216 instances.

MFRC The Moral Foundation Reddit Corpus (Trager et al., 2022) consists of 16k Reddit posts, annotated according to the updated Moral Foundation Theory (Atari et al., 2023) (**fairness** is split into two separate foundations: *equality* and *proportionality*). In addition, the authors introduce the category *thin morality* in order to anticipate instances of weak moral signal that would naturally lead to a higher disagreement. Similarly to Hoover et al. (2020), the annotators were encouraged to be consistent while relying on their intuition instead of forcing a high level of agreement. However, the final pool of annotators consists of only 5 who were chosen based on reliability and availability.

After the same pre-processing has been applied as before, the dataset contains 17,457 instances, each being annotated by between three and five annotators. For the majority vote we assign *thin morality* for all cases in which there was no clear majority vote (2851 instances). In cases of a tie between a foundation and *non-morality* we choose the foundation as the majority vote.

Human Values For Human Values we rely on the Touché23-ValueEval dataset that was also used as part of a SemEval Shared Task in 2023 (Kiesel et al., 2022, 2023). The dataset contains 9k arguments sourced from diverse domains (religious texts, newspaper editorials, free-text written arguments), each argument consists of a premise that either supports or rejects a conclusion. For annotating human values (Kiesel et al., 2022) developed a multilayer taxonomy in which the most fine-grained layer consists of 54 values. The presence or absence of each value is annotated in a binary fashion, and an annotator has to annotate all 54 values for a given text. This makes the annotation process on the one hand cumbersome; on the other hand, fine-grained values are more concrete and specific and therefore easier to annotate (e.g. *be polite, be honest, have harmony with nature*. The most specific level of this multi-layer scheme was thus used for annotation to achieve a higher level of agreement.

Individual values are then mapped to a more coarse-grained schema that consists of 20 value categories that are used as target labels for the Shared Task.

The final dataset consists of 9,233 arguments annotated by 38 different annotators. Each instance is annotated by three or four annotators. The majority vote for this dataset was a bit easier as the label *no value* does not exist. In cases of a tie between the absence or presence of a specific value, we choose the presence as majority vote.

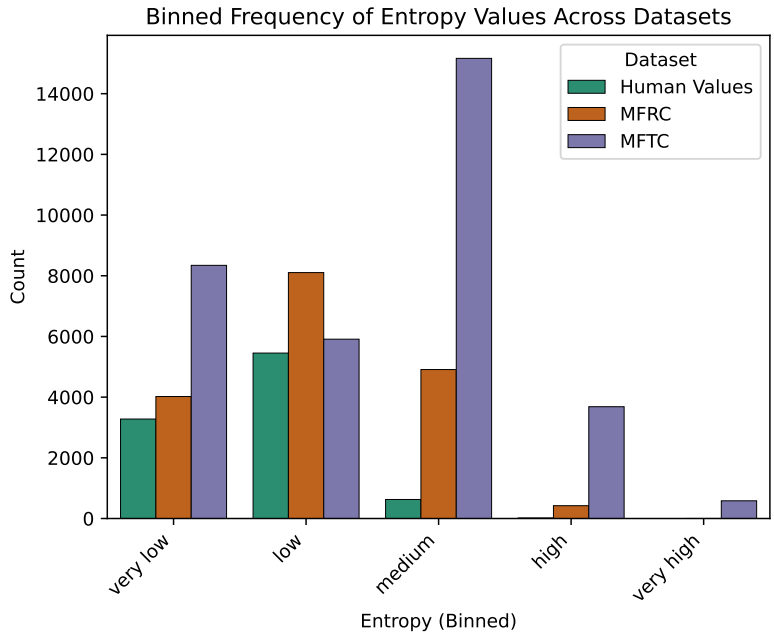

Figure 6: Distribution over different entropy levels (amount of HLV) – comparison between the three datasets.

A.1 Distribution of disagreement

B Linguistic features: details

B.1 Feature extraction and validation

For the complexity and polarity features, as well as for the number of named entities we use a dedicated Python package (Maurer, 2025).

For the pragmatic features, we rely on pre-trained classifiers. The storytelling classifier (`falkne/storytelling-LM-europarl-mixed-en` has been developed by (Falk and Lapesa, 2022) and is trained on three different sources of argumentation data, one of which is Reddit, thereby matching at least two of our domains. We use the classifier that resulted in the most robust performance across different domains to increase out-domain generalizability.

For irony (Barbieri et al., 2020) (`cardiffnlp/twitter-RoBERTa-base-irony`) and sarcasm (`mrm8488/t5-base-finetuned-sarcasm-twitter` we use classifiers that were developed for twitter, thereby matching at least on of our domains.

We acknowledge that all features were extracted automatically, making them susceptible to some degree of error. However, re-annotating and manually validating a large set of features is beyond the scope of this work. If a linguistic feature contains noise, that noise is expected to be consistent across all items within a dataset. With an interpretation based on regression analysis, which accounts for the combined effects of all features, significant effects are assumed to emerge beyond the noise. Since analyses are conducted separately for each dataset, any unreliability in a feature would apply consistently across an entire dataset.

C Regression analysis: procedure

In the different analyses presented in the paper, we follow the regression analysis setup outlined below.

We fit and inspect linear regression models with linguistic features as predictors (independent variable) and human label variation (section 4.2) or model uncertainty (section 5.3) as predicted values (dependent variable).

As a preliminary step, we carry out feature selection based on the feature correlations and their explanatory power. We do this before the regression analysis to avoid multicollinearities. Given that we extracted in total more than around 160 features, we extract the 30 most explanatory features for each dataset in a first step, using each feature as a single independent variable.

Feature	Feature name	Description	Feature Type
Number of sentences	n_sentences	Number of sentences in the text	Complexity
Average word length	avg_word_length	Average word length (in characters): n_characters / n_tokens	Complexity
Lexical density	lexical_density	Lexical density of the text: n_lexical_tokens / n_tokens	Complexity
Flesch reading ease	flesch_reading_ease	Flesch reading ease score of the text	Complexity
Average concreteness	avg_concreteness	Average human concreteness ratings of the tokens in the text	Complexity
Average prevalence	avg_prevalence	Average human prevalence ratings of the tokens in the text	Complexity
Average valence	avg_valence	Average valence of the tokens in the text	Polarity
Average dominance	avg_dominance	Average dominance of the tokens in the text	Polarity
Average emotion intensity for anger	avg_intensity_anger	Average intensity of an emotion	Polarity
Average emotion intensity for surprise	avg_intensity_surprise	Average intensity of an emotion	Polarity
Number of hedge words	n_hedges	Number of hedge words in the text (words expressing uncertainty of the speaker).	Pragmatic
Storytelling	storytelling	Presence of a personal experience or narrative	Pragmatic
Sarcasm	sarcasm_score	Probability that a text contains sarcastic language	Pragmatic
Irony	irony	presence of irony in a text	Pragmatic
Number of named entities	n_entities	Number of named entities in the text	Lexical
Number of named entities of type numeral	n_num	Number of named entities in the text with type numerals	Lexical
Number of loaded words	count_values	Number of words from moral foundation or value dictionary that are present in the text.	Lexical

Table 4: List of significant features (total number of features that were found to be significant for any type of regression model, both in HLV and in model uncertainty) with their description.

	count	mean	std	min	25%	50%	75%	max	dataset
lexical_density	9376.000	0.499	0.105	0.111	0.429	0.500	0.560	1.000	HV
avg_intensity_trust	9376.000	0.370	0.262	0.000	0.000	0.469	0.567	0.844	HV
avg_word_length	9376.000	4.696	0.679	2.667	4.233	4.624	5.063	10.250	HV
storytelling	9376.000	0.005	0.069	0.000	0.000	0.000	0.000	1.000	HV
avg_valence	9376.000	0.592	0.098	0.175	0.530	0.601	0.660	0.958	HV
n_sentences	9376.000	1.258	0.691	1.000	1.000	1.000	1.000	9.000	HV
avg_concreteness	9376.000	2.501	0.234	1.613	2.348	2.491	2.640	3.996	HV
count_values	9376.000	3.080	2.294	0.000	1.000	3.000	4.000	22.000	HV
avg_dominance	9376.000	0.574	0.062	0.318	0.533	0.575	0.615	0.800	HV
n_hedges	9376.000	1.359	1.319	0.000	0.000	1.000	2.000	12.000	HV
n_entities	9376.000	0.385	1.029	0.000	0.000	0.000	0.000	19.000	HV
n_num	9376.000	0.076	0.331	0.000	0.000	0.000	0.000	6.000	HV
flesch_reading_ease	9376.000	-2051.795	947.281	-11039.535	-2566.644	-1886.165	-1335.250	-93.832	HV
avg_prevalence	9376.000	2.304	0.055	1.825	2.279	2.311	2.338	2.516	HV
avg_intensity_anger	9376.000	0.210	0.276	0.000	0.000	0.000	0.469	0.959	HV
avg_intensity_surprise	9376.000	0.091	0.185	0.000	0.000	0.000	0.000	0.930	HV
lexical_density	33684.000	0.394	0.143	0.000	0.304	0.389	0.478	1.000	MFTC
avg_intensity_trust	33684.000	0.317	0.306	0.000	0.000	0.406	0.609	0.844	MFTC
avg_word_length	33684.000	5.297	2.024	1.000	4.214	4.920	5.864	113.375	MFTC
storytelling	33684.000	0.049	0.216	0.000	0.000	0.000	0.000	1.000	MFTC
avg_valence	33684.000	0.535	0.195	0.000	0.437	0.554	0.665	1.000	MFTC
n_sentences	33684.000	1.724	0.977	1.000	1.000	1.000	2.000	14.000	MFTC
avg_concreteness	33684.000	2.529	0.592	0.000	2.297	2.552	2.818	5.000	MFTC
count_values	33684.000	1.923	1.810	0.000	1.000	2.000	3.000	21.000	MFTC
avg_dominance	33684.000	0.522	0.146	0.000	0.468	0.538	0.602	0.991	MFTC
n_hedges	33684.000	0.548	0.902	0.000	0.000	0.000	1.000	10.000	MFTC
n_entities	33684.000	1.796	1.602	0.000	1.000	1.000	3.000	25.000	MFTC
n_num	33684.000	0.220	0.594	0.000	0.000	0.000	0.000	19.000	MFTC
flesch_reading_ease	33684.000	-1036.443	770.842	-6511.075	-1497.345	-824.605	-476.055	205.820	MFTC
avg_prevalence	33684.000	2.205	0.377	-0.842	2.217	2.288	2.334	2.576	MFTC
avg_intensity_anger	33684.000	0.263	0.313	0.000	0.000	0.000	0.561	0.964	MFTC
avg_intensity_surprise	33684.000	0.080	0.180	0.000	0.000	0.000	0.000	0.930	MFTC
lexical_density	17457.000	0.391	0.084	0.000	0.338	0.390	0.443	0.833	MFRC
avg_intensity_trust	17457.000	0.395	0.252	0.000	0.000	0.484	0.570	0.867	MFRC
avg_word_length	17457.000	4.006	0.722	1.778	3.636	3.944	4.280	31.067	MFRC
storytelling	17457.000	0.153	0.360	0.000	0.000	0.000	0.000	1.000	MFRC
avg_valence	17457.000	0.588	0.082	0.000	0.542	0.591	0.638	0.958	MFRC
n_sentences	17457.000	2.695	1.663	1.000	2.000	2.000	3.000	20.000	MFRC
avg_concreteness	17457.000	2.510	0.233	0.000	2.368	2.500	2.636	4.850	MFRC
count_values	17457.000	1.988	2.056	0.000	0.000	1.000	3.000	16.000	MFRC
avg_dominance	17457.000	0.544	0.063	0.000	0.507	0.545	0.583	0.814	MFRC
n_hedges	17457.000	2.376	2.424	0.000	1.000	2.000	3.000	20.000	MFRC
n_entities	17457.000	1.881	2.248	0.000	0.000	1.000	3.000	21.000	MFRC
n_num	17457.000	0.307	0.838	0.000	0.000	0.000	0.000	23.000	MFRC
flesch_reading_ease	17457.000	-1455.501	912.733	-12001.600	-1840.820	-1249.635	-861.322	133.628	MFRC
avg_prevalence	17457.000	2.276	0.105	-0.055	2.250	2.295	2.325	2.507	MFRC
avg_intensity_anger	17457.000	0.253	0.283	0.000	0.000	0.074	0.500	0.959	MFRC
avg_intensity_surprise	17457.000	0.144	0.207	0.000	0.000	0.000	0.258	0.930	MFRC

Table 5: Descriptive statistics of all pre-selected features

We then applied hierarchical clustering to those pre-selected features for each dataset and manually selected them based on distinct clusters and whether there was overlap between datasets. This resulted a set of 24 features balanced between explanatory power and overlap between datasets. The final set of features (Table 4 gives an overview of them and their description, and Table 5 reports their descriptive statistics) amounts to 17 in total, because these were found to be significant in any of the final models.

We then proceed to linear regression with step-wise selection (R package StepAIC, standard settings), considering simple effects. We do not consider interactions at this stage because this would add a complexity level in the interpretation that would exceed the scope of this paper. Once step-wise selection has identified the best models (i.e., the most predictive set of features) we proceed to inspect the results, in three steps:

- The **amount of explained variance of the model** (R^2) indicates the extent to which our features, as a whole, impact the distribution of the predicted values of interest (human or model uncertainty). In other words: how well can we predict variance in uncertainty at a global level?
- The **amount of explained variance of specific feature groups** (R^2) indicates the amount of variance in the predicted values that can be ascribed to them: i.e., do complexity features or pragmatic features impact human disagreement the most?
- The analysis of the **estimates of the significant predictors** builds the ground for our interpretation of individual effects: i.e., does higher complexity lead to higher disagreement or uncertainty?

D Regression analysis on annotator disagreement

D.1 Explained variance

Feature type	RoBERTa-mv	RoBERTa-KLd	flan-t5-crowd	flan-t5-MV	flan-t5-RL
Complexity	4.741	8.525	3.54	6.459	3.661
Lexical	3.269	3.717	0.697	3.214	0.437
Polarity	4.611	8.17	3.25	7.873	0.427
Pragmatic	0.523	0.803	0.434	0.763	0.039
Total	13.145	21.215	7.92	18.308	4.564

Table 6: Regression analysis of model uncertainty, MFRC (moral foundations, Reddit): total amount of explained variance stemming from the linguistic factors for each of the three datasets. For each feature category: the sum of explained variance for features that belong to that category.

Feature type	RoBERTa-mv	RoBERTa-KLd	flan-t5-crowd	flan-t5-MV	flan-t5-RL
Complexity	7.918	11.69	10.916	3.088	0.052
Lexical	1.912	1.437	2.531	0.232	0.013
Polarity	3.306	2.56	3.818	0.68	0.016
Pragmatic	1.463	1.933	0.527	0.409	0.013
Total	14.599	17.62	17.792	4.409	0.095

Table 7: Regression analysis of model uncertainty, MFTC (moral foundations, Twitter): total amount of explained variance stemming from the linguistic factors for each of the three datasets. For each feature category: the sum of explained variance for features that belong to that category.

D.2 Estimate plots: impact of linguistic features on model uncertainty

D.3 Examples of items with high disagreement and corresponding features

Table 9 displays examples with high disagreement along with the most explanatory features identified in our analysis.

Feature type	RoBERTa-mv	RoBERTa-KLd	flan-t5-crowd	flan-t5-MV	flan-t5-RL
Complexity	5.286	2.36	2.454	7.353	0.818
Lexical	3.338	0.281	0.188	6.484	0.249
Polarity	2.726	3.882	2.209	2.59	3.787
Pragmatic	0.3	–	–	0.133	0.138
Total	11.65	6.523	4.851	16.56	4.992

Table 8: Regression analysis of model uncertainty, Human Values : total amount of explained variance stemming from the linguistic factors for each of the three datasets. For each feature category: the sum of explained variance for features that belong to that category.

text	entropy	feature value	feature	corpus
polarity				
Every man needs to remind @realDonaldTrump what respect and integrity mean. Humility and compassion are necessary to serve as a leader. #NotmyPresident #impeachtrump	0.61	0.70	avg_valence	MFTC
Look at all these dumb, scumbag, redneck, racist, misogynist white trash. Lol" "HOW COULD THEY EVER VOTE FOR TRUMP THEY DESTROYED AMERICA WHEN IT WAS SO WONDERFUL (FOR ME)"	0.57	0.50	avg_dominance	MFRC
some people fear of retribution from those that oppose their thoughts and ideas. this is a place they can feel safe.	0.50	0.62	avg_dominance	HV
pragmatic				
In 99% of professional jobs, no company would hold your hand and sacrifice for the employee if they were struggling and hurting the team. It's basically just yet another way capital owners try to keep workers from making things difficult for them. Loyalty to your employer in our modern economy sounds like such a feudal idea.	0.57	1	irony	MFRC
#BLM has put my own nervousness around cops in perspective. Now I'm like, calm down woman, the worst they're going to do is ticket you.	0.24	1	storytelling	MFTC
complexity				
#BlackLivesMatter = demand for their humanity #AllLivesMatter = denial of systematic racism #BlueLivesMatter = refusing accountability	0.61	0.65	lexical density	MFTC
Self-determination is a basic collective human right. Catalans practiced this right in 2017. Since then the Catalan Republic is kidnaped by Spanish lawfare, with the elected president Puigdemont in exile within Europe and more than 3000 Catalan activists suffering political persecution as a consequence. Ignoring this conflict within the core of Europe by European Institutions can undermine the EU at its very start. If you sincerely believe in human rights, recognize the Catalan Republic officially. This crucial step will enforce European democracy, ...	0.27	6	n_sentences	HV
Legalizing marijuana would lead to a reduction in gang-related drug violence.	0.0	2.87	concreteness	HV
lexical				
#BBCWorld #TheresaMay #Foxnews #TrumpTariffs Turns out defending #DonaldTrump for 2years is excellent practice for defending putting kids in cages[...]	0.60	15	named entities	MFTC
I do not care at all about people stealing from big ass corporations who do not treat employees fairly and barely pay them a livable wage. These companies have insurance theyll be alright	0.573	5	nr. of loaded words	MFRC

Table 9: Examples of the three datasets with entropy (human label variation) and the value of one characteristic feature.

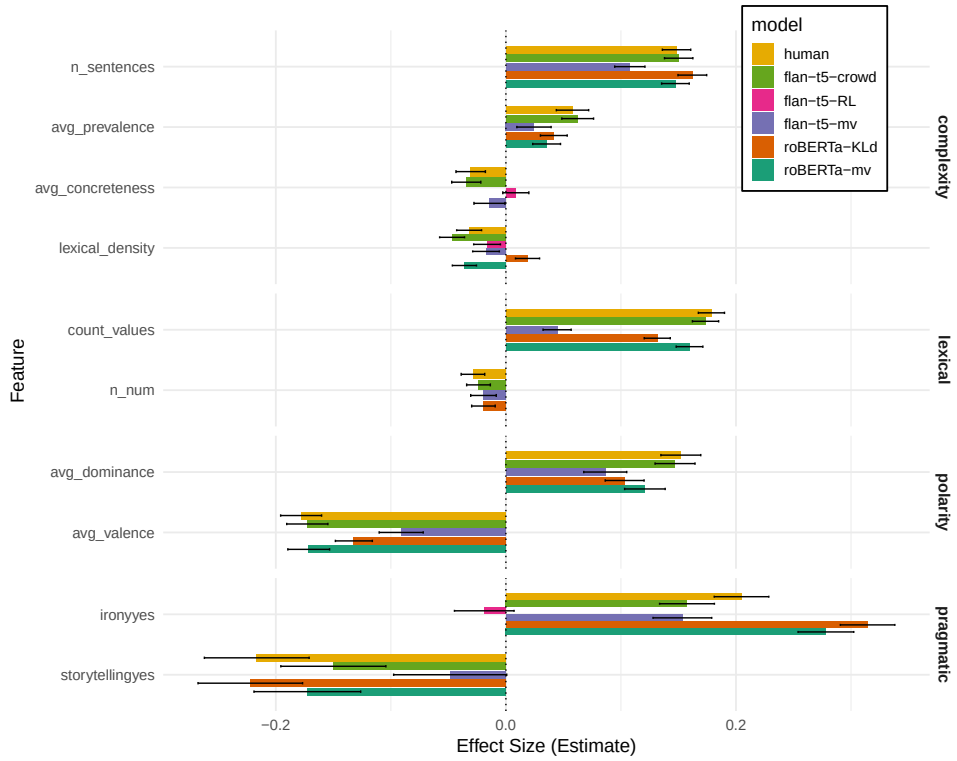

Figure 7: MFTC corpus (moral foundations, Twitter) – regression analysis of model uncertainty: estimates and direction of effects for the most explanatory models selected by AIC, grouped by feature type

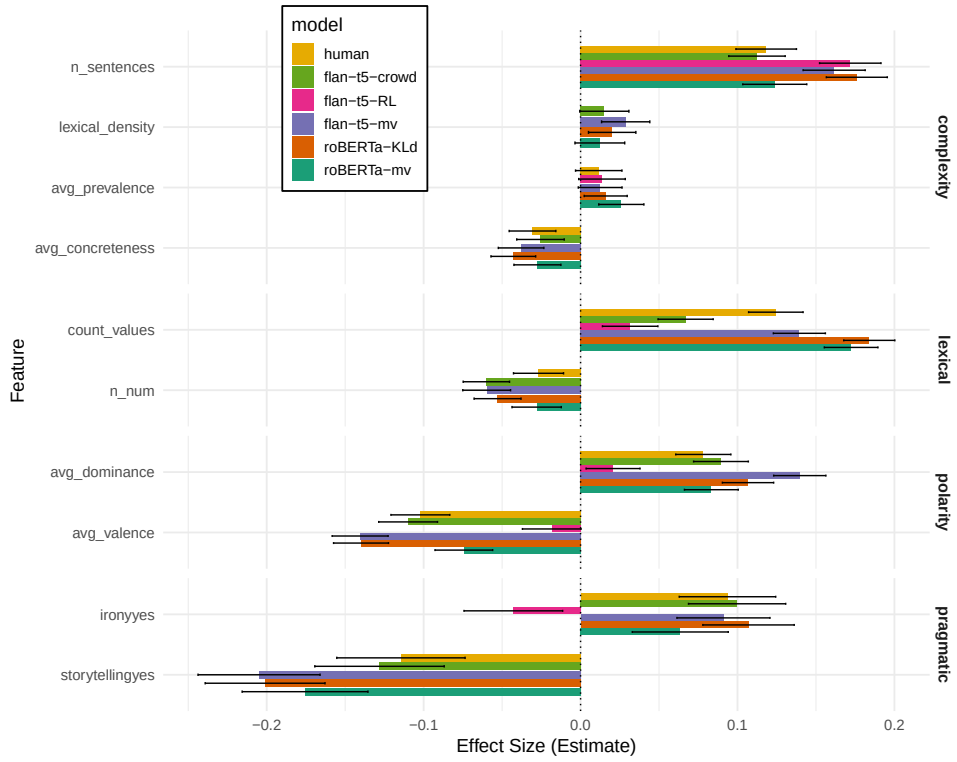

Figure 8: MFRC corpus (moral foundations, Reddit): regression analysis of model uncertainty: estimates and direction of effects for the most explanatory models selected by AIC, grouped by feature type

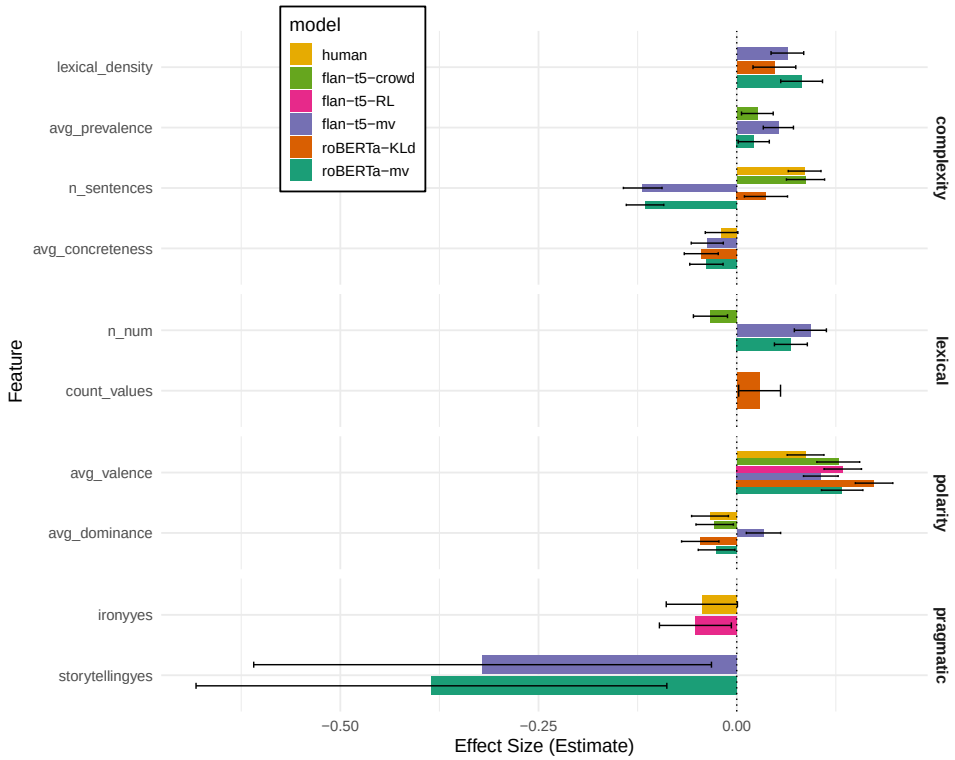

Figure 9: Human values corpus – regression analysis of model uncertainty: estimates and direction of effects for the most explanatory models selected by AIC, grouped by feature type

E Implementation and training details

We first divide the human average disagreement (as measure by the mean over normalized entropy scores for each label) into five equal-width intervals to have different levels of agreement (from low to high). We use a stratified Kfold split (n=5) and stratify based on agreement level, such that instance of different agreement levels are equally represented in training, validation and test. The size of the splits is 60 percent for training, 20 percent for test and validation.

We train RoBERTa-base (both, the majority vote-based and the KLd-based) for 20 epochs and use the model that received the highest macro-F1 score on the validation set as the best model. We apply class weights to counteract class imbalance. The model has 125M parameters.

We use LoRA as a parameter-efficient finetuning (PEFT) method to instruction fine-tune the flan-t5-large model. The model has a total of 792.6 million parameters, of which only 9.4 million parameters are trainable. This means that approximately 1.19% of the model's parameters are being updated during training, while the remaining parameters remain fixed. We train all flan-based models for 4 epochs (independent of the dataset).

E.1 Instructions for the LLM and further post-processing.

We use create two different instructions. Instruction a. is used for flan-t5-mv and flan-t5-RL. Instruction b. is used for flan-t5-crowd.

a. `Instruction: You are given an argumentative text and a predefined set of categories indexed in a label2index dictionary. Your task is to evaluate each label and determine whether it applies (1) or does not apply (0) to the text. Output Format: Return a binary vector where the i-th position corresponds to the label in the label2index dictionary. Steps: Analyze the text to identify relevant information. For each label, make a binary decision: 1: If the label applies to the text. 0: If the label does not apply. Return the binary vector in the same order as the label2index dictionary. Remember: Base your decisions solely on the input text. Align the binary vector strictly with the label2index ordering. label2index: non-moral: 0, care: 1, fairness: 2, loyalty: 3, authority: 4, purity: 5`

b. `Instruction: You are given an argumentative text and a predefined set of categories indexed in a label2index dictionary. Your task is to evaluate each label and determine how many out of 30 annotators agree with the label. Output Format: Return a numerical vector where the i-th position corresponds to the label in the label2index dictionary, and the value represents the number of annotators that agree with the label (between 0 and 30). Steps: Analyze the text to identify relevant information. For each label, count how many out of 30 annotators agree with the label. Return the numerical vector in the same order as the label2index dictionary. Remember: Base your decisions solely on the input text. Align the numerical vector strictly with the label2index ordering. label2index: non-moral: 0, care: 1, fairness: 2, loyalty: 3, authority: 4, purity: 5`

We experimented with different formats as well (e.g., using a likert scale to indicate agreement as in Juroš et al. (2024)) but opted for the two methods reported in the paper, since it was more straightforward to retrieve a fine-grained probability distribution from them which is comparable to the other models.

Note that in some cases the generated output is not valid (e.g. wrong format, wrong label size). This sometimes reduces the sample size from 30 to a lower one. However, we find that most generated instances were valid.

F Experimental results: calibration to human label variation

Figure 3 visualizes the density curves for the TVD across all instances, comparing low versus high disagreement.

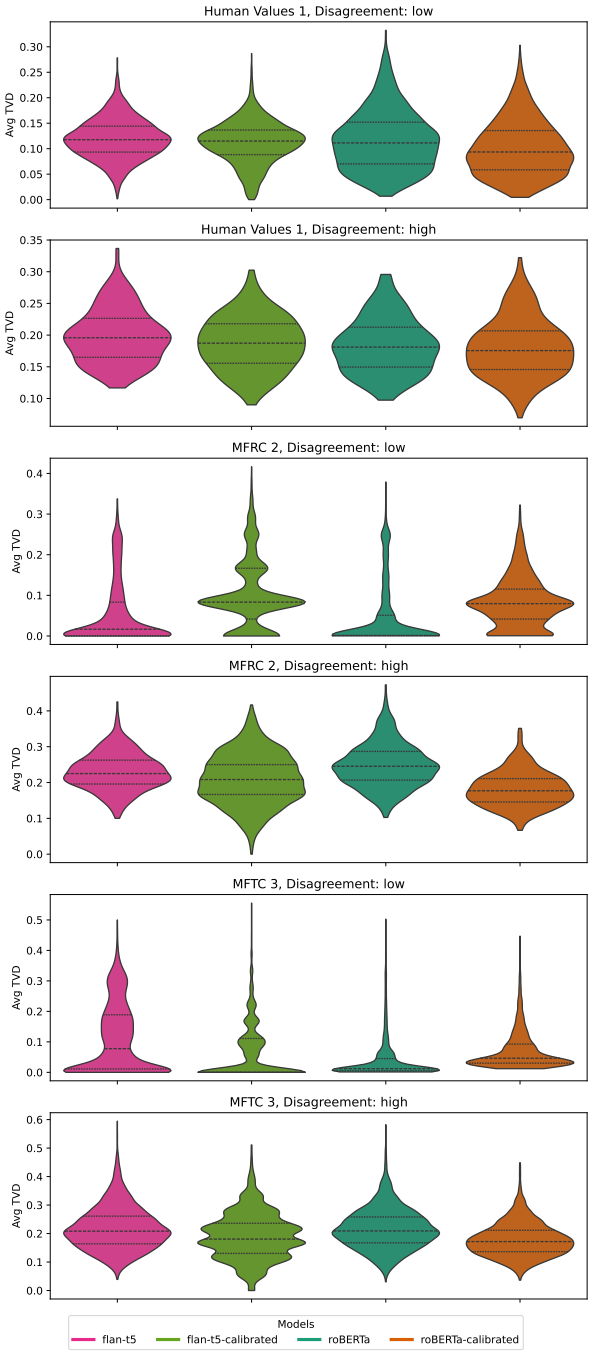

Figure 10: Density plot for the instance-based TVD metric. Comparing the different models for low-disagreement and high disagreement data.

G Generative Assistance in Authorship

We used generative AI for assistance purely with the language of the paper (paraphrasing or polishing the author's original content). We used generative AI as coding support (github co-pilot) when programming the scripts for the experiments conducted in this paper.