# Moral Disagreement over Serious Matters: Discovering the Knowledge Hidden in the Perspectives

**Anny Álvarez, Oscar Araque**
Universidad Poltécnica de Madrid, ETSI Telecomunicación, Intelligent Systems Group
Avenida Complutense, 30, Madrid, 28040, Spain
a.anogales@alumnos.upm.es, o.araque@upm.es

### Abstract

Moral values significantly define decision-making processes, notably on contentious issues like global warming. The Moral Foundations Theory (MFT) delineates morality and aims to reconcile moral expressions across cultures, yet different interpretations arise, posing challenges for computational modeling. This paper addresses the need to incorporate diverse moral perspectives into the learning systems used to estimate morality in text. To do so, it explores how training language models with varied annotator perspectives affects the performance of the learners. Building on top if this, this work also proposes an ensemble method that exploits the diverse perspectives of annotators to construct a more robust moral estimation model. Additionally, we investigate the automated identification of texts that pose annotation challenges, enhancing the understanding of linguistic cues towards annotator disagreement. To evaluate the proposed models we use the Moral Foundations Twitter Corpus (MFTC), a resource that is currently the reference for modeling moral values in computational social sciences. We observe that incorporating the diverse perspectives of annotators into an ensemble model benefits the learning process, showing large improvements in the classification performance. Finally, the results also indicate that instances that convey strong moral meaning are more challenging to annotate.

**Keywords:** moral foundations theory, language models, perspectivism

## 1. Introduction

The language we use mirrors our thoughts, emotions, values, and cultural background, shaping our interactions with others. The proliferation of online communication platforms and social media has empowered individuals to voice and disseminate their opinions on contentious issues rapidly and to a larger audience. Under these circumstances it is relevant to assess the attitude of individuals towards certain topics of interest. Moral values play an essential role in shaping our decision-making process, particularly when addressing contentious subjects. When dealing with issues such as global warming or political regulations, individuals reference their moral value system, consciously or subconsciously. The Moral Foundations Theory (MFT) has been developed to interpret the concept of morality across diverse cultures (Haidt and Joseph, 2004), outlining five core foundations: *care, fairness, loyalty, authority*, and *sanctity*. The MFT has benefited from refinement with the addition of a sixth foundation: *liberty* (Haidt, 2012).

Despite being recent, the MFT is currently a well-established theory in psychology and the social sciences. Besides, it has found broad acceptance in the field of computational social science due to the creation of a clear taxonomy of values and the development of several computational resources, such as the Moral Foundations Dictionary (MFD) (Graham et al., 2009), which serves as a central resource for natural language processing applications. The creators of the MFD report some challenges involved in the construction process of such a resource since linguistic, cultural and historical contexts influence language usage.

Attending to the nature of moral values, the MFT has been designed with the idea of harmonizing the variety of moral expressions across different cultures. That is, the MFT models innate foundations that are common to different cultures. Of course, this also means that different cultures and thus, individuals will instantiate the moral foundations differently under the same circumstances. This shows one of the key challenges of generating computing models of the MFT: considering different moral perspectives on the same topic.

While the current datasets and lexicons (Hoover et al., 2020; Trager et al., 2022) do consider the annotations of different individuals, ultimately these annotations are treated in an aggregated manner (i.e., using a voting mechanism) and do not explore the richness introduced by a diverse set of annotators. This lack of understanding of morality computational models introduces a severe bias that can influence individuals (Krügel et al., 2023). Moreover, recent works highlight the necessity of considering a diverse set of annotations simultaneously, without recurring to aggregations that lose relevant information (Cabitza et al., 2023). In light of this, this work explores the information contained within a set of annotators when modeling morality in an attempt to shed light on such a relevant issue.

Thus, we explore the effect of considering the views from several annotators in an already annotated moral dataset, the Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020). In doing so, this paper investigates the impact of training different language models with the perspective of each annotator and then combining these models in an ensemble fashion. Additionally, the task of assessing whether an instance is particularly challenging to annotate is considered, providing further insight into the language usage of this type of text.

To frame the contributions of the paper, we explore the following research questions (RQs). **RQ1: To what extent can the diversity of views in moral annotations be useful for automated moral assessments?** This work examines the variance of the annotations of the MFTC, training different language models with different annotations. Using these trained models, we explore the effect of this additional knowledge in the framework of automatically estimating morality in text.

Following, we also inspect **RQ2: Is it possible to automatically assess whether a text is challenging to annotate?** This question reflects on the characteristics of texts where annotators diverge in their ratings, offering a basis on which we can understand the difficulties of evaluating moral foundations. In this sense, this paper evaluates the performance of several models in the task of predicting whether a text is challenging to annotate, using the disagreement that the annotators of the MFTC have shown.

The rest of the paper is structured as follows. Section 2 describes the fundamentals of the Moral Foundations Theory (MFT) and how it has been previously addressed from a computational perspective. Section 3 presents the data and methodology used in this work. Next, the experimentation is detailed in Section 4. Finally, the conclusion and future work is delineated in Section 5.

## 2. Background

In this section, we summarize key concepts and methodologies for our research. First, we explore the Moral Foundations Theory (MFT), which represents the underlying principles that influence human moral judgments in diverse cultural contexts and resources such as the Moral Foundations Dictionary (MFD). We also discuss the application of these resources in computational models, including the use of prompts, which has demonstrated the potential to enhance the comprehension and generation of texts.

### 2.1. Moral Foundations Theory

Previously, it has been mentioned that the Moral Foundations Theory (MFT) describes, through the definition of several foundations, common axes to measure morality across diverse cultures and sensibilities. In this work, we study the five basic foundations (Haidt and Joseph, 2004). *Care/harm*: This foundation relates to our capacity to empathize with and perceive the pain of others. It encompasses virtues such as kindness, gentleness, and nurturance. *Fairness/cheating*: This foundation underscores the virtues of justice and rights. *Royalty/betrayal*: manifests the principles of solidarity. It embodies virtues like patriotism and willingness for group-oriented self-sacrifice. *Authority/subversion*: This foundation emphasizes virtues associated with leadership and followership. It entails deference to esteemed authority figures and reverence for traditional norms. *Purity/degradation*: This foundation emphasizes aspirations for elevated living, often found in religious narratives. It encompasses virtues of self-discipline, self-improvement, naturalness, and spirituality.

We have already covered that one of the main reasons the MFT has become so popular in computational social sciences is the development of the Moral Foundations Dictionary (MFD) (Graham et al., 2009). This lexical resource, based on the known Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), covers a basic annotation of lemmas and how they convey meanings toward the moral foundations. While this resource has been crucial for the development of computational models of morality, it is not without limitations. Among the notable limitations of the MFD are: (i) a limited number of tokens; (ii) inclusion of “radical” lemmas seldom encountered in everyday language, such as “homologous” and “apostasy”; and (iii) classification based on a moral bipolar scale denoting vice and virtue, lacking any indication of “strength.”

Concerning the dataset we use in this work, the Moral Foundations Twitter Corpus has been used in several scientific studies and natural processing tasks whose work is based on the MFT (Graham et al., 2013) and the MFTC as a reference to evaluate distinct moral narratives in natural language texts. On one hand, this was used to study how a moral lexicon (Araque et al., 2020) can be exploited at the document level using different machine learning and engineering techniques, obtaining better results in the detection of morality in text. On the other hand, Guo et al. (2023) propose a refinement model that uses Sentence-BERT embeddings to capture moral information, investigating the performance, generalisation and transferability of moral embeddings with a specific focus on how these embeddings can improve the accuracy of moral classifiers. Finally, Liscio et al. (2022) perform an extensive investigation on the effects of cross-classification of moral values in text, comparing a deep learning model on seven different domains.

### 2.2. Prompts for inserting knowledge

The utility of pre-trained language models for a large variety of natural language processing applications is clear due to its success and popularity (Han et al., 2021). In this regard, language models show characteristics in their internal representations and behaviors that indicate that are they capable of generating a depiction of moral concepts (Scherrer et al., 2023; Fitz, 2023). For example, it has been found that language models' internal representations induce a moral dimension that, in principle, could be utilized by the model (Fitz, 2023). We argue that this kind of morality knowledge can be exploited to assess moral values in text.

Following on the previous, one common method to control the output of a language model is to steer their generation process through *prompts*. Prompts are instructions or fragments designed to guide the model during the performance of a specific task. Although this approach has not been previously used in the context of moral values assessment, we build on the evidence of positive results obtained in other tasks using pre-trained models such as BERT (Luo et al., 2022).

For an comprehensive review on the use of prompts, please consult the work of Liu et al. (2023).

## 3. Data and Methods

As described, this work pursues to gain insights into how an already annotated dataset can be used to characterize different perspectives in the process of annotating moral values in text. This section describes the dataset used in the experimentation (Sect. 3.1) and the methods designed to explore the knowledge of the annotations (Sect. 3.2).

### 3.1. Dataset

To perform the experiments detailed in Section 4, we have used the Moral Foundations Twitter Corpus dataset (Hoover et al., 2020) and its corresponding annotations. It is structured into seven subsets of data, each addressing distinct and socially relevant discursive topics. The corpus has been labelled by various annotators; it is composed of a considerable size of tweets (approximately 35 thousand ) and a diversity of ideas in various social movements, from politics and human rights to natural disasters. These aspects provide a comprehensive view of how morality is reflected in different social media, thus making it a benchmark for machine learning tasks such as multi-labelled morals.

Originally, the dataset consists of seven different subsets that contain Twitter messages pertaining to different societal issues: All Lives Matter(ALM), Black Lives Matter (BLM), Baltimore, Davidson, Election, MeToo Movement (MT) and Sandy. We work with 6 of them, which are available online[1]. These are the following: All Lives Matter (ALM), related to 'All Lives Matter' Movement; Black Lives Matter (BLM), related to 'Black Lives Matter' Movement; Baltimore, related to the Baltimore protest following the death of Freddie Gray in US; Davidson, texts collected by Davidson et al. (2017) for hate speech and offensive language research; Election, tweets about the 2016 US presidential election; and Sandy, related to Hurricane Sandy in 2012.

This set of human-annotated English tweets has labels of moral foundations in 10 classes distinguishing between vice and virtue for each moral trait, including a 'non-moral' class. Tweets were tagged following the MFT, described in Section 2.1, and each domain was evaluated by at least three trained annotators as set out in the original labeling guide (Hoover et al., 2017), which has been designed as a comprehensive manual that establishes common practices and clear guidelines for the identification of moral sentiments expressed in texts. Despite the training given to annotators, the authors put emphasis on the use of personal views even if they diverged from common values, increasing the variety in the annotations. Each tweet was therefore labelled with an indication of the presence or absence of each virtue and vice or using a 'non-moral' label.

In this study, a basic pre-processing and subsequent tokenization has been carried out to the data, as required by this type of transformer model. Numbers, punctuation marks, symbols, usernames, URLs, and emoticons were removed, and stopwords were preserved. The final label for each text was obtained by aggregating the labels of several annotators using the majority vote as the true class, resulting in the distribution of morality found in each dataset and reflected in Table 1.

To assess the overview of different annotators, we set each annotator's label to the corresponding text the person had annotated. Table 5 shows the final distribution of labels per annotator.

One observable concern is the imbalance towards the 'non-moral' class, where in Davidson and Baltimore cases, they are approximately 90% of the total. Although we use the original data to take advantage of the largest dataset, these limitations were taken into account when analyzing the results and reflecting on the conclusion.

[1] https://osf.io/k5n7y/

| Dataset | C/H | F/C | L/B | A/S | P/D | NM |
|---|---|---|---|---|---|---|
| ALM | 1,314 | 723 | 408 | 274 | 182 | 585 |
| BLM | 1,048 | 934 | 528 | 491 | 253 | 1,040 |
| Baltimore | 434 | 292 | 895 | 120 | 37 | 2,366 |
| Davidson | 447 | 130 | 319 | 1,039 | 118 | 2,784 |
| Election | 798 | 736 | 286 | 177 | 349 | 2,019 |
| Sandy | 708 | 708 | 1,010 | 519 | 560 | 291 |

Table 1: Distribution of foundations presence in all data domains. The column names are encoded as follows. C/H: care/harm, F/C: fairness/cheating, L/B: loyalty/betrayal, A/S: authority/subversion, P/D: purity/subversion, NM: non-moral.

## 3.2. Methodology

To satisfy the research questions previously raised (see Sect. 1), this work studies (i) how the information of the disagreement among annotators can be exploited, as well as (ii) the characteristics of what constitutes an instance prone to be subject to disagreement.

For all experiments, we have used Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018)[2] as the base model. Given the unbalance of the data labels and its effect on the performance in the classification tasks, all results are reported using the macro-averaged F-score.

Regarding the model specifications, it was used the pre-trained BERT model `bert-base-uncased` along with its corresponding tokenizer. Each model was trained for 15 epochs, using a batch size of 32 and learning rates of 0.01 and 2e5 respectively.

**Diversity exploitation.** Regarding the first challenge, this work proposes an evaluation that probes the utility of understanding the moral views of the different annotators. In this regard, we first assess the variety of the annotators by training a model that predicts the moral of the text as judged by each annotator. As Figure 1 illustrates, we fine-tune a different instance of the same model using as training labels the annotations expressed by each annotator. In this way, we intend that each captures the particularities and views of each annotator. Additionally, we evaluate the classification performance of each of these models, which can offer further insights into the consistency of the annotations.

Following, a supplemental evaluation is done. To predict the aggregated label of each data instance, we use the previously fine-tuned models trained on the specific annotations of each annotator and the corresponding text.

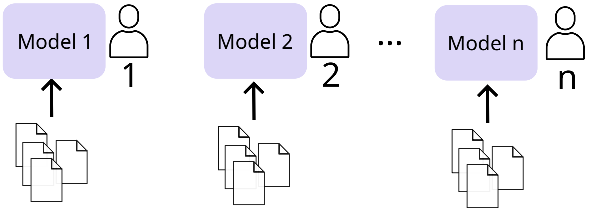

Figure 1: Fine-tuning procedure where models are trained with the specific annotations of $n$ different annotators.

To carry out this evaluation, we decided to explore the use of a prompt-based approach adding the predictions of each fine-tuned model as additional information alongside the original text. Then, a second training was performed using the enriched dataset to analyse how it contributed to the performance of the model in the classification task. This is shown in Figure 2.

The choice of this strategy is based on the proven effectiveness of these models in natural language understanding and leveraging the ability to capture semantics, incorporating multiple perspectives through the predictions of morals provided by different annotators. We believe that this approach could provide a more complete and refined view of the moral dimensions present in the data, which in turn could improve the performance of models on the moral classification task.

Feeding the model in a consistent way with the perspectives of each annotator enriches the dataset by providing it with additional information about each text, especially about the different model perspectives it may contain. Taking into account the limitation of choosing the prompt template manually due to the numerous possibilities and choosing the one that maximises the performance of the model, a structure has been used that reflects as clearly as possible that the additional information conveys the view of different annotators.

During the evaluation of diversity explained above, the predictions of each model were used for each data instance and annotator. These predictions were added to the standardised prompt at the input, following the structure: *'The text {...} has been annotated by different annotators with the following moral values {* $m_1, m_2, \dots m_n$ *}'*, where {...} is the original input and { $m_1, m_2, \dots m_n$ } is a concatenation of the annotations for the text.

By providing these, we can better align the predictions with the characteristics and evaluation features of each text, improving the accuracy and consistency in the prediction of the aggregated labels. Once the new inputs were obtained, the training of the BERT model was performed, and the results were compared with the base training.

Thereby, we propose that having an overview

[2] https://huggingface.co/google-bert/bert-base-uncased

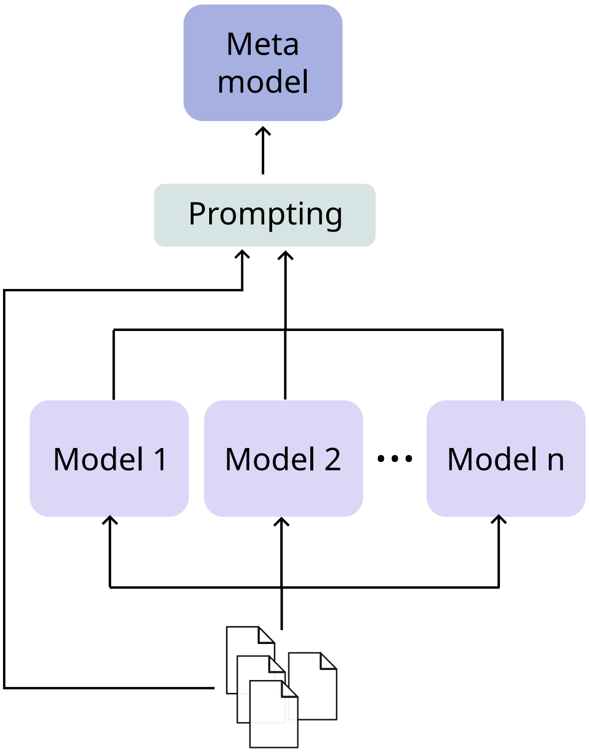

Figure 2: Proposed ensemble method that combined the predictions on the different perspectives of the base models with the textual intput through a prompt approach.

of each of the annotator's subjective perspectives can aid in the overall estimation of morality in text. To assess this, this work compares the ensemble method to the baseline of estimating morality using solely the text. These models, as shown in Figures 1 and 2, are thoroughly evaluated in Section 4.1.

**Disagreement estimation.** To address the second research question, we propose the use of a learning model to assess whether a text is challenging to annotate. This task is oriented to exploit the information inherent in the disagreement among annotators, where some instances will show a high agreement and other instances' content will be harder to annotate. This proposition follows the ideas presented by (Basile, 2020) in the sense that it is an attempt to consider all different perspectives contained within the original annotations in a machine learning setting.

In this way, we fine-tune a different instance of a BERT model for each dataset, having as label the level of disagreement of the data instance. To facilitate the analysis, we have considered a binary approach so that each instance can be considered as either challenging to annotate (i.e., that shows a high disagreement among annotators) or not. Thus, in this proposal, the learning models perform a binary classification task: given a document, predict whether the instance is challenging to annotate.

To assess if a given instance is positive or negative under the mentioned distinction, we define a *divergence* metric that allows us to encode this idea of agreement among annotators. More formally, consider a set of annotations for a given data instance $A = \{a_1, a_2 \cdots a_N\}$; we then define a measure of agreement among annotators. Thus, the agreement for annotator $i$ is defined as:

$$g_i = \frac{1}{N} \sum_{i \neq j} a_i == a_j \tag{1}$$

where $N$ is the number of annotations. The $==$ operation returns a value of $1$ if $a_i = a_j$, and a value $0$ otherwise. Naturally, $g_i$ encodes the number of times that annotator $i$ agrees with the rest of the annotators for that data instance. Thus, $g = \{g_1, g_2, \cdots, g_N\}$ Following, we define the divergence metric as the opposite of the previous:

$$d = 1 - \frac{g}{\max(g)} \tag{2}$$

where $d \in [0, 1]$. The closer $d$ is to $0$, the less divergent the instance (i.e., the more agreement among annotators); conversely, the closer $d$ is to $1$, the higher disagreement in the annotations we observe. Therefore, we utilize the divergence metric $d$ as a measure to identify instances that are challenging to annotate.

Since we are modeling the problem through a binary approach, a threshold concerning the divergence metric has been defined. Thus, we consider an instance to be challenging to annotate if $d \geq d_{th}$. Section 4.2 describes how this threshold has been estimated.

Finally, to study the characteristics of the language in documents that have diverging annotations, we use the SHapley Additive exPlanations (SHAP) method (Lundberg and Lee, 2017). Such a method assigns an importance score to each of the features considered for an specific prediction. These SHAP values allow us to inspect the learning models trained, inspecting how the language affects the decision on the disagreement of a document.

To perform this analysis, we extract the SHAP values of all models trained, aggregating them to obtain a whole overview of the classification process. To do so, we extract the SHAP values for all words in all documents, aggregating them into a set of values for each word considered.

These evaluations, which address the estimation of disagreement, are described in Section 4.2s.

# 4. Experimentation

In this section, we present the results obtained. Concretely, Section 4.1 focuses on both the individual performance of the fine-tuned models according to different annotators and the impact of using these predictions as additional knowledge for morality prediction. Following, Section 4.1 describes the analysis done on the modeling of agreement among annotators.

## 4.1. Annotation diversity exploitation

Firstly, Table 2 presents the results of the performance evaluation of the models on each dataset and for each annotator, comparing them to the baseline results.

| Dataset | Annot. | Baseline | F1-Score |
|---|---|---|---|
| ALM | 00 | 64.71 | 13.46 (-51.24) |
| | 01 | | 60.52 (-04.18) |
| | 02 | | 26.22 (-38.48) |
| | 03 | | 79.35 (+14.64) |
| BLM | 00 | 85.46 | 79.44 (-06.01) |
| | 01 | | 79.38 (-06.07) |
| | 02 | | 37.94 (-47.51) |
| | 03 | | 83.05 (-02.40) |
| | 04 | | 83.55 (-01.90) |
| Baltimore | 02 | 42.58 | 40.17 (-02.40) |
| | 13 | | 39.59 (-02.98) |
| | 14 | | 49.54 (+06.96) |
| Davidson | 05 | 15.84 | 15.21 (-00.62) |
| | 06 | | 15.50 (-00.33) |
| | 07 | | 14.64 (-01.19) |
| Election | 00 | 61.11 | 58.40 (-02.70) |
| | 02 | | 32.24 (-28.86) |
| | 03 | | 65.57 (04.46) |
| | 04 | | 70.81 (09.70) |
| Sandy | 09 | 55.73 | 56.58 (00.85) |
| | 10 | | 52.73 (-02.99) |
| | 11 | | 48.49 (-07.23) |

Table 2: Results of the classification performance in predicting the moral as judged by the different annotators.

It can be observed that there is significant variability in performance across different datasets and between different annotators. One relevant observation is that better results are found in the cases where there is a more balanced distribution of classes. Additionally, we argue that the interpretation of moral values may depend significantly on the context and domain of the text, which can influence the consistency and accuracy of different annotator's labels.

In general, the results only diverge slightly from the baseline results, except for annotator 00 in the ALM dataset, where it performs much worse. The pronounced disparities observed in some cases are mainly due to the amount of data labelled by these annotators. An insufficient number of examples prevents the model from accurately learning and predicting the labels assigned by these annotators.

The lowest metric values are observed in the Davidson dataset. This is likely due to class imbalance and subjectivity in the interpretation of moral values in this specific context. In the Davidson case, approximately 60% of the labelled data was identified as 'non-moral'. For more details on the class distributions for each annotator, see Table 5.

Finally, as reflected in Table 3, in terms of the model's performance when using prompts, a significant improvement in the classification performance was obtained in all domains compared to the baseline model without prompts. This suggests that the choice of prompt and additional information on different perspectives can influence and improve the results.

The incorporation of this additional information has effectively provided more contextual cues, allowing the model to better understand and classify morality in different texts across various domains. Moreover, the observed improvements in F1 scores highlight the effectiveness of leveraging diverse perspectives from annotators. By adding these into the training process, the model becomes more efficient at recognizing moral nuances present in texts. However, it's remarkable that while the prompt-based approach has led to considerable enhancements, certain domains, like Davidson, still present challenges for accurate classification. Overall, the success of using prompts underscores the significance of contextual information and diverse perspectives in morality estimation tasks.

| | F1-score | |
|---|---|---|
| | Baseline | Prompting |
| ALM | 64.71 | 88.74 |
| BLM | 85.46 | 95.82 |
| Baltimore | 42.58 | 76.32 |
| Davidson | 15.84 | 66.03 |
| Election | 61.11 | 88.22 |
| Sandy | 55.73 | 86.44 |

Table 3: Evaluation of the addition of different perspectives in training. The F1-Score results are compared with baseline results in all domains.

## 4.2. Disagreement estimation

As described in Section 3.2, we study the nature of the disagreement among annotators by training a learning model to predict whether a given text is challenging to annotate. By approaching the issue in this manner, we are operating on the basis that annotators diverge in their annotations driven by certain characteristics of the texts they are annotating.

Firstly, we have defined a threshold $d_{th}$ on the divergence metric that allows us to distinguish whether a text is challenging to annotate. Figure 3 shows the evolution of the percentage of positive instances, that is, instances that show a divergence metric where $d > d_{th}$. Based on the distributions of the $d$ metric along all datasets, we manually set this threshold to $d_{th} = 0.7$. As can be seen, the majority of the distributions in the figure suffer an abrupt decline when the threshold is at the indicated number.

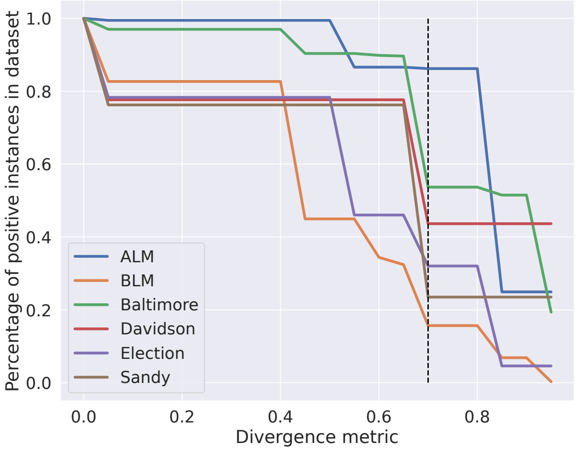

Figure 3: Percentage of instances considered to be challenging (vertical axis) to annotate with the divergence metric (horizontal axis).

Following this, fine-tuning and evaluation of the learning models has been performed. We have trained a different instance of the same model for each of the data domains in an attempt to capture the specific characteristics of each domain. To avoid the negative effect of the imbalance of the two classes considered, we balanced the resulting data by randomly sampling the majority class. The results of such an experiment are shown in Table 4, including the number of derived instances for each data domain.

It is clear, attending to the results, that in some cases the classifier is able to distinguish the divergent instances (i.e., the instances where annotators show a higher divergence metric). These cases include the BLM, Baltimore, and Election domains, with the highest performance metrics. In contrast, in the ALM, Davidson, and Sandy domains, the classifiers are not able to properly discern the divergence of the data instances, although for ALM and Sandy the f-score reaches 58%. This dissimilar behaviour among domains is a consistent result: as studied previously by Liscio et al. (2023), the differences in the domains of the Moral Foundations Twitter Corpus (MFTC) do affect the quality of prediction tasks.

| | Acc. | F1-score | Neg. inst. | Pos. inst. |
|---|---|---|---|---|
| ALM | 58.55 | 58.36 | 94 | 99 |
| BLM | 68.94 | 68.49 | 120 | 115 |
| Baltimore | 79.26 | 79.25 | 402 | 355 |
| Davidson | 48.17 | 47.75 | 442 | 403 |
| Election | 71.61 | 71.39 | 272 | 288 |
| Sandy | 58.82 | 58.81 | 180 | 177 |

Table 4: Evaluation in the task of predicting whether a text is challenging to annotate with morality. Accuracy, macro averaged F-score, and the number of negative and positive instances are reported.

Overall, these positive results are a clear indication that there are language cues that indicate to the learners whether a text is prone to be challenging to annotate. Since these language signals are sure to vary with the domain of annotations, we seek to gain a better understanding of this process. To do so, as described in Section 3.2, we use SHAP to inspect how the learners analyse the text in terms of divergent annotations. In this study, we have aggregated the SHAP values from all data domains, as we aim to obtain a general view of this process rather than a specific examination of each domain's particularities.

Figure 4 shows the results obtained from a selection of the tokens that have the highest relevance for either the negative or positive classes. Tokens with negative SHAP values are relevant for detecting the negative class (i.e., instances that show low disagreement), while tokens with positive SHAP values are related to detecting the positive class, where the disagreement is higher.

We observe that the tokens with negative SHAP values are generally words with semantics not pertinent to morality and innocuous in terms of societal or cultural issues. Interesting examples of these terms are *photo*, *wonderful*, *green*, *internet* or *babies*. This is an intuitive result since annotators will generally agree within texts that do not convey a strong moral or cultural position. In contrast, tokens with positive SHAP values tend to express strong moral significance. Some examples of these words are *democrats*, *evil*, *god*, *duty*, *racism*, *homo* (from homosexuality), and *respect*. Again, this can be explained if we consider that annotators will disagree more frequently when assessing documents that include morally and culturally stronger positions. Interestingly, some tokens with higher positive SHAP values revolve around polemic or even harmful matters such as religion, sexual practices, and racism.

To better understand the insights obtained by this last study, we include some interesting examples of texts that show the characteristics found through the SHAP analysis. For instance, the following text, contained in the All Lives Matter (ALM) dataset, "*'#blacklivesmatter is for unity equality respect between races all lives matter ignores the truth of injustice to claim reverse racism'*" has been annotated with the foundations care, loyalty and fairness by the different annotators, which indicate that the annotators have identified different foundations in the text, although all of them are virtues as defined in the MFT.

Another instance, extracted from the Sandy domain, is as follows: "*Sandy is god's way of saying ignoring climate change is equal to saying you are willing to destroy my creation*". This text has been annotated with the foundations of authority, purity, and fairness. While the purity and authority annotations probably reference the religious content, the debatable fairness annotation may relate to a sense of divine justice, alluded to in the original message.

## 5. Conclusion

This paper explores the effect of diverse human annotations in the context of computationally modeling moral foundations through the Moral Foundations Theory. Under the lenses of perspectivism (Cabitza et al., 2023) [3], we explore a known dataset in the field of moral value estimation, the Moral Foundations Twitter Corpus. This dataset contains annotations from different annotators that are commonly aggregated. This work investigates the effect of separately considering the perspectives of the annotators toward morality.

Concretely, we raise two research questions (RQs) that are thoroughly studied in this work. Firstly, RQ1 inspects the effect of exploiting the diversity of annotators' perspectives for automated moral estimation. In this regard, we have shown that the different annotators do highly impact the quality of the predictions if taken in isolation. Attending to this, it is clear that the diversity of annotators and domains are variables to take into account when generating new data repositories. In contrast, the experiments show notable and consistent improvements in the classification performance when adding the predictions of models trained to estimate individual annotators' perspectives into an ensemble model. Such a positive result motivates future research on harnessing diverse perspectives into learning systems.

Secondly, RQ2 proposes the task of estimating whether a data instance is challenging to annotate. That is, if an instance generates disagreement among annotators. Through this task, we intend to analyse the linguistic cues that indicate disagreement factors. The experiments show that the ability to estimate disagreement can achieve high performance scores but varies across domains, indicating considerable variance. By doing a subsequent analysis using SHAP values, we have discovered that the disagreement instances tend to contain strong moral, political, or cultural meanings. On the contrary, instances where annotators typically agree normally contain more neutral language.

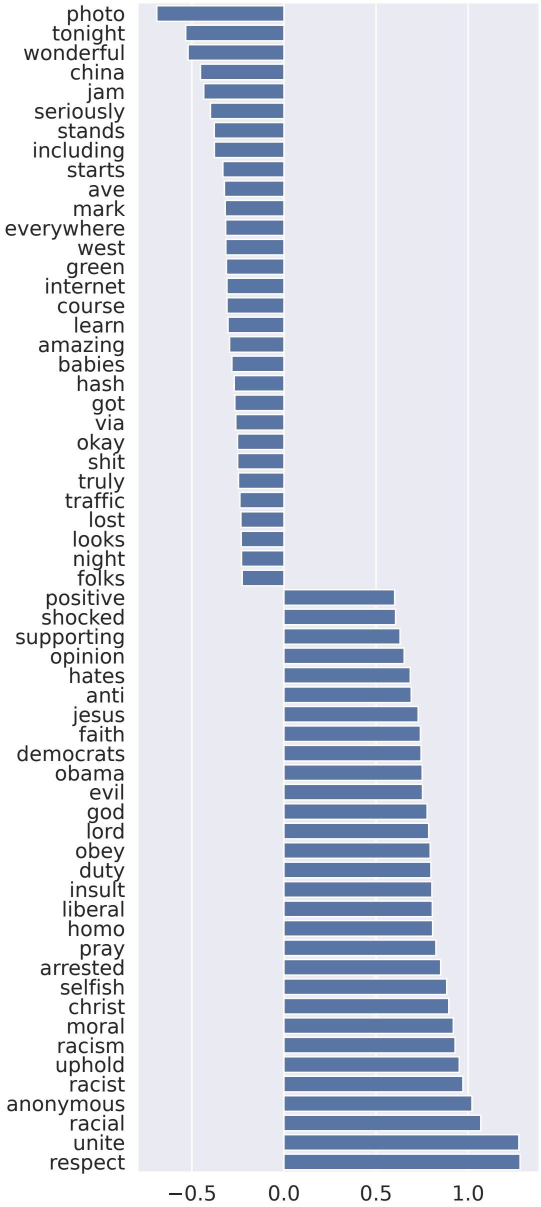

Figure 4: SHAP values of interesting tokens. Positive values indicate relevance towards the positive class, while negative values indicate otherwise.

Addressing the limitations of the work, we evaluate the ensemble method using an aggregated label for moral values. Oddly, this challenges one of the principles of the perspectivism movement,

[3] The perspectivist data manifesto: https://pdai.info/.

which states that traditional *golden* labels should be avoided, thus taking into account the diversity of views from annotators. This part of the proposed evaluation does simplify the challenge of moral estimation for the ensemble method due to the large complexity involved in designing a model that predicts over such a substantial set of target labels (i.e., all possible combinations of moral foundations for each of the annotators). Future work should tackle this issue by modeling the prediction objective more tractable.

Another limitation of the work is related to our definition of what constitutes a divergent instance. We have defined a straightforward metric that aids in defining a learning problem related to disagreement. In this regard, future work should investigate this direction, further defining the divergence of annotated documents and how we can handle them.

## 6. Acknowledgements

This project has been funded by the project UNICO I+D Cloud - AMOR, financed by Ministry of Economic Affairs and Digital Transformation, and the European Union through Next Generation EU.

## 7. Appendix

Table 5 describes the distributions of moral annotations for each annotator and data domain.

After a first analysis of the different annotations in the texts, it was observed that there was a disparity in the amount of data labelled by each annotator. In order to ensure a correct comparison, we initially used the intersection of instances annotated by all the annotators. However, this strategy faced the challenge of dealing with very small datasets due to annotators with minimal contributions. Thus, to overcome this problem four annotators from different domains were removed.

For ALM dataset, annotator00 was excluded because only 94 instances were labelled, which is a considerable lower proportion in comparison to the 3486 instances from the original dataset. In the case of Baltimore dataset, annotator12 and annotator15 were also discarded for their low contribution. Finally, in Davidson dataset, annotator08 was removed because their annotations consisted in 1 instance.

Removing these annotators was done to prevent the datasets from being too small and negatively impacting the training process. In the case of Baltimore dataset, when all the annotators were considered, the data was reduced from 4144 examples to 402, resulting in significant missing data and poor metrics in performance. Excluding annotator 12 and 15 results in a large dataset formed by the intersection of 3528 examples, leading to better model performance.

## 8. Bibliographical References

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Stephen Fitz. 2023. Do large gpt models discover moral dimensions in language representations? a topological study of sentence embeddings. *arXiv preprint arXiv:2309.09397*.

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.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 281–291.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

| Dataset | Annotator | C/H | F/C | L/B | A/S | P/D | NM |
|---|---|---|---|---|---|---|---|
| ALM | 00 | 48 | 26 | 7 | 4 | 4 | 5 |
| | 01 | 1252 | 714 | 403 | 266 | 181 | 524 |
| | 02 | 1299 | 718 | 404 | 273 | 179 | 579 |
| | 03 | 1312 | 722 | 408 | 273 | 182 | 582 |
| BLM | 00 | 881 | 774 | 448 | 435 | 222 | 858 |
| | 01 | 906 | 784 | 458 | 441 | 236 | 856 |
| | 02 | 910 | 787 | 452 | 443 | 228 | 870 |
| | 03 | 924 | 795 | 457 | 446 | 236 | 874 |
| | 04 | 923 | 793 | 456 | 447 | 234 | 864 |
| Baltimore | 02 | 390 | 267 | 806 | 115 | 35 | 2054 |
| | 12 | 102 | 84 | 130 | 33 | 4 | 599 |
| | 13 | 423 | 281 | 855 | 113 | 34 | 2313 |
| | 14 | 426 | 286 | 884 | 114 | 36 | 2280 |
| | 15 | 61 | 60 | 100 | 30 | 3 | 276 |
| Davidson | 05 | 435 | 113 | 290 | 957 | 118 | 2705 |
| | 06 | 396 | 121 | 294 | 1036 | 95 | 2121 |
| | 07 | 150 | 38 | 161 | 174 | 74 | 1880 |
| | 08 | 1 | 0 | 0 | 0 | 0 | 0 |
| Election | 00 | 728 | 667 | 263 | 161 | 335 | 1991 |
| | 02 | 736 | 673 | 265 | 165 | 327 | 1922 |
| | 03 | 791 | 733 | 283 | 175 | 345 | 2004 |
| | 04 | 791 | 734 | 285 | 177 | 347 | 1951 |
| Sandy | 09 | 701 | 702 | 990 | 517 | 554 | 285 |
| | 10 | 706 | 701 | 989 | 514 | 553 | 289 |
| | 11 | 703 | 705 | 992 | 509 | 540 | 290 |

Table 5: Distribution of foundations by annotator. The column names are encoded as follows. C/H: care/harm, F/C: fairness/cheating, L/B: loyalty/betrayal, A/S: authority/subversion, P/D: purity/subversion, NM: non-moral.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Xianchang Luo, Yinxing Xue, Zhenchang Xing, and Jiamou Sun. 2022. Prcbert: Prompt learning for requirement classification using bert-based pretrained language models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24 – 54.

## 9. Language Resource References

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.

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.

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.

Joseph Hoover, Kate Johnson-Grey, Morteza Dehghani, and Jesse Graham. 2017. Moral values coding guide. *PsyArXiv*.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.