What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric

Liscio, E.; Araque, Oscar; Gatti, Lorenzo; Constantinescu, I.L.; Jonker, C.M.; Kalimeri, Kyriaki; Murukannaiah, P.K.

DOI
10.18653/v1/2023.acl-long.789

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers

Citation (APA)
Liscio, E., Araque, O., Gatti, L., Constantinescu, I. L., Jonker, C. M., Kalimeri, K., & Murukannaiah, P. K. (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* (pp. 14113–14132). (Proceedings of the Annual Meeting of the Association for Computational Linguistics; Vol. 1). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2023.acl-long.789

What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric

Enrico Liscio[1], **Oscar Araque**[2], **Lorenzo Gatti**[3], **Ionut Constantinescu**[4], **Catholijn M. Jonker**[1,6], **Kyriaki Kalimeri**[5], and **Pradeep K. Murukannaiah**[1]
[1]TU Delft, Delft, the Netherlands [2]Universidad Politécnica de Madrid, Madrid, Spain
[3]University of Twente, Enschede, the Netherlands [4]ETH Zürich, Zürich, Switzerland
[5]ISI Foundation, Turin, Italy [6]Leiden University, Leiden, the Netherlands
{e.liscio,c.m.jonker,p.k.murukannaiah}@tudelft.nl o.araque@upm.es
l.gatti@utwente.nl iconstantinescu100@gmail.com kyriaki.kalimeri@isi.it

Abstract

Moral rhetoric influences our judgement. Although social scientists recognize moral expression as domain specific, there are no systematic methods for analyzing whether a text classifier learns the domain-specific expression of moral language or not. We propose Tomea, a method to compare a supervised classifier's representation of moral rhetoric across domains. Tomea enables quantitative and qualitative comparisons of moral rhetoric via an interpretable exploration of similarities and differences across moral concepts and domains. We apply Tomea on moral narratives in thirty-five thousand tweets from seven domains. We extensively evaluate the method via a crowd study, a series of cross-domain moral classification comparisons, and a qualitative analysis of cross-domain moral expression.

1 Introduction

Moral narratives play a fundamental role in stance taken on controversial social issues (Fulgoni et al., 2016). Recognizing moral narratives helps understand the argumentation around important topics such as vaccine hesitancy (Kalimeri et al., 2019b), violent protests (Mooijman et al., 2018), and climate change (Dickinson et al., 2016).

Language reveals deep psychological constructs, including moral values (Graham et al., 2013). Thus, language is an important avenue for analyzing moral expression. In particular, supervised text classification models have been showing promising results on morality prediction (Lourie et al., 2021; Hendrycks et al., 2021; Alshomary et al., 2022). These models leverage the wisdom of crowds (via annotations of moral expression) to attain a descriptive understanding of morality. However, the supervised learning paradigm can lead to black-box models (Danilevsky et al., 2020). Understanding what these models learn is crucial, especially for the morality classification task, which is likely to be used in sensitive applications like healthcare (Wen et al., 2019; Carriere et al., 2021).

Moral expression is *context* dependent (Hill and Lapsley, 2009; Brännmark, 2015; Kola et al., 2022), where context refers to factors such as actors, actions, judges, and values (Schein, 2020). For a text classifier, the *domain* from which the training data is sourced represents the context. For example, in the context of recent Iranian protests, tweets tagged *#mahsaamini* can form the training domain. We expect this domain to have a different moral expression than the training domain of *#prolife* tweets, representing a different context.

Recent works (Liscio et al., 2022a; Huang et al., 2022) analyze the out-of-domain performance of morality classifiers. However, what leads classifiers to perform differently across domains has not been systematically explored. Such an insight is essential for understanding whether classifiers can learn a domain-specific representation of morality.

We propose Tomea (from the Greek $\tau o \mu \acute{\epsilon} \alpha$, meaning "domain") to compare a text classifier's representation of morality across domains. Tomea employs the SHAP method (Lundberg and Lee, 2017) to compile domain-specific *moral lexicons*, composed of the lemmas that the classifier deems most predictive of a moral concept in a domain, for each moral concept and domain. Through such moral lexicons, Tomea enables a direct comparison of the linguistic cues that a classification model prioritizes for morality prediction across domains.

We employ Tomea to compare moral rhetoric across the seven social domains in the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020). Then, we perform a crowdsourced evaluation to assess the agreement between the human intuition and the automatically obtained results of Tomea. We show that this agreement is consistent across domains but varies across moral concepts. Further, we find a strong correlation between the results of Tomea and the out-of-domain performance

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14113–14132

of the models used for obtaining the moral lexicons. In addition, we perform qualitative analyses of the moral impact of specific lemmas, unveiling insightful differences in moral concepts and domains.

Tomea allows to inspect and compare the extent to which a supervised classifier can learn domain-specific moral rhetoric from crowdsourced annotations. Tomea can guide computer scientists and practitioners (e.g., social scientists or policymakers) in the responsible use of transfer learning approaches. In transfer learning, large datasets are used to pre-train language models, which are then finetuned with data collected in the domain of interest. Such pre-training typically helps in improving performance in the finetuning domain. However, increased performance may come at the cost of critical mistakes which may hinder the usage of the model, especially when the finetuning domain concerns minority groups (Nadeem et al., 2021). Tomea can assist in the qualitative comparison of pre-training and finetuning domains by unveiling potential critical differences and guiding practitioners in judging the appropriateness of using a morality prediction model in an application.

2 Related Works

We introduce the theoretical background and review related works in morality classification in text, domain dependency in NLP models, and explainability in NLP.

Moral Theories The expression of morality in language has been explored via constructs such as rules-of-thumb on acceptable social behavior (Forbes et al., 2020), moral norms (Lourie et al., 2021; Emelin et al., 2021), and ethical judgements (Hendrycks et al., 2021). However, these constructs are too abstract for our purpose of understanding the domain-specific expression of morality.

We base our work on models of *human values*, which represent morality in the form of innate moral elements. Two well-known models of human values are the Moral Foundation Theory (MFT) (Graham et al., 2013) and the Schwartz Theory of Basic Human Values (Schwartz, 2012).

In this work, we explore the domain-specific expression of moral elements of the MFT. The MFT consists of five foundations, each consisting of a vice–virtue duality, resulting in 10 moral elements, as shown in Table 1. We choose the MFT because of the availability of the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020), a corpus of seven datasets corresponding to seven domains (Section 4.1), enabling cross-domain analyses.

Element	Definition
Care/ Harm	Support for care for others/ Refrain from harming others
Fairness/ Cheating	Support for fairness and equality/ Refrain from cheating or exploiting others
Loyalty/ Betrayal	Support for prioritizing one's inner circle/ Refrain from betraying the inner circle
Authority/ Subversion	Support for respecting authority and tradition/ Refrain from subverting authority or tradition
Purity/ Degradation	Support for the purity of sacred entities/ Refrain from corrupting such entities

Table 1: The moral elements (virtue/vice) of MFT.

Morality Classification Classification of moral elements in text has been approached via moral lexicons, lists of words depictive of moral elements. Lexicons are generated manually (Graham et al., 2009; Schwartz, 2012), via semi-automated methods (Wilson et al., 2018; Araque et al., 2020), or expanding a seed list with NLP techniques (Ponizovskiy et al., 2020; Araque et al., 2022). The lexicons are then used to classify morality using text similarity (Bahgat et al., 2020; Pavan et al., 2020). Moral elements have also been described as knowledge graphs to perform zero-shot classification (Asprino et al., 2022).

More recent methods adopt instead supervised machine learning (Qiu et al., 2022; Alshomary et al., 2022; Kiesel et al., 2022; Liscio et al., 2022a; Huang et al., 2022; Lan and Paraboni, 2022). A textual dataset is annotated with the moral elements, and the resulting labels are used to train a supervised model. This approach represents the starting point for our analysis in this paper.

Domain Dependency Domain dependency is a well-known issue in sentiment analysis (Al-Moslmi et al., 2017), where it is often addressed through domain adaptation, the challenge to adapt a lexicon or a machine learning algorithm to a novel domain (Hamilton et al., 2016; Wu and Huang, 2016; Wilson and Cook, 2020; Mohamad Beigi and Moattar, 2021). Our main goal in this paper is to analyze the differences in morality across domains, but not to adapt a lexicon or a model to novel domains.

Explainability Explainable AI (XAI) has been used extensively in NLP (Danilevsky et al., 2020).

We do not contribute a new method to XAI, but our work is a novel application of an XAI method.

A key distinction is whether an XAI method generates local or global explanations. Local explanations expose the rationale behind an individual prediction, e.g., by highlighting the most important words in a sentence (Ribeiro et al., 2016; Lundberg and Lee, 2017). Global explanations expose the rationale behind the whole decision-making of the model, e.g., by inducing taxonomies of words that are predictive of the classified labels (Pryzant et al., 2018; Liu et al., 2018). In our analysis, we induce lexicons to explain the decision-making of the models, as they provide an intuitive global explanation.

3 The Tomea Method

Tomea[1] is a method for comparing a text classifier's representation of morality across domains. Tomea takes as input two ⟨dataset, classifier⟩ pairs, where, in each pair, the classifier is trained on the corresponding dataset. Since Tomea intends to compare moral expressions across domains, the two datasets input to it are assumed to be collected in different domains. Tomea's output is a qualitative and quantitative representation of the differences in moral expressions between the two input domains.

Figure 1 shows the two key steps in the method. First, we generate *moral lexicons* capturing the classifiers' interpretable representations of the moral elements specific to their domains. Then, we compare the moral lexicons in two ways. (1) We compare the moral lexicons generated for the same moral elements in different domains. (2) We combine the moral lexicons generated for the same domains and provide a single measure of moral rhetoric similarity between two domains.

3.1 Moral and Domain Lexicons

A *moral lexicon* represents how a morality classifier interprets the expression of a moral element in a domain. We represent the expression of morality by determining the impact that each word has toward the classification of a moral element in a domain. Thus, a moral lexicon consists of (w, i) pairs, where w in each pair is a word that the classifier considers relevant for predicting the examined moral element in the domain under analysis and i is its impact. This way, we generate a lexicon for each moral element in each domain. We refer to

[1] https://github.com/enricoliscio/tomea

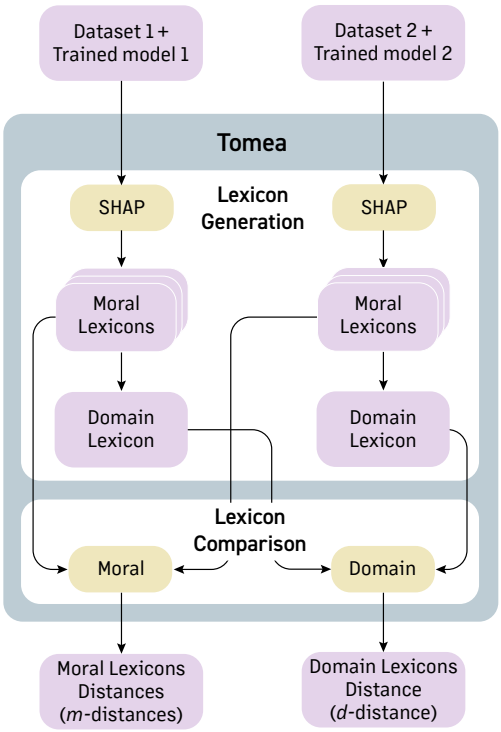

Figure 1: Tomea takes as input two ⟨dataset, model⟩ pairs (where the datasets are collected in different domains) and returns the distance in moral expressions across moral elements and domains.

the union of the moral lexicons generated for all moral elements in a domain as the *domain lexicon*.

3.2 Lexicon Generation

We use Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) to generate the lexicons. SHAP uses Shapley values to quantify the extent to which an input component (a word) contributes toward predicting a label (a moral element).

The impact of a word is computed as the *marginal contribution* of the word toward a label prediction. Intuitively, the marginal contribution of the word is calculated by removing the word from the sentence and evaluating the difference between the sentence with and without the word. All combinations of words in the sentence (i.e., the *power set* of features) are created to compute the impact of each word. The resulting impact is positive (if the likelihood of predicting a certain label increases when the word is present) or negative (if the likelihood decreases). We aggregate the local explanations to obtain a global ranking of word impact for each moral element. This can be done by adding the local impact of words for each entry of the dataset due to the additive nature of SHAP.

Tomea executes the following steps to obtain

moral lexicons from a dataset and a model. (1) Execute SHAP on each entry of the dataset with the related model, resulting in a (w, i) pair for each word that appears in the dataset. (2) Replace each word w with its lemma, if one can be found using NLTK's WordNet-based lemmatizer (Bird et al., 2009). (3) Combine words that share the same lemma by adding their impact i together.

3.3 Lexicon Comparison

Tomea enables the comparisons of (1) moral lexicons across domains, and (2) domain lexicons.

Moral Lexicons First, Tomea normalizes each moral lexicon by substituting each word's impact with its z-score (Triola, 2017) based on the distribution of the impact scores of all words in a moral lexicon. Then, Tomea computes an *m*-distance (moral element distance) to compare the lexicons of a moral element generated in different domains.

Let $W = \{w_1, \cdots, w_n\}$ be the set of n common words between the moral lexicons of a moral element M_i (one of the ten in MFT) in the two domains D_A and D_B (in practice, all words that appear in both lexicons). Then, let the two vectors,

$$\mathbf{i}^{(D_A,M_i)} = [i_1^{(D_A)}, \cdots, i_n^{(D_A)}] \text{ and}$$
$$\mathbf{i}^{(D_B,M_i)} = [i_1^{(D_B)}, \cdots, i_n^{(D_B)}],$$

represent the impacts of the words belonging to W on M_i in domains D_A and D_B, respectively.

Then, the *m*-distance compares the impacts that the same set of words has in the two domains D_A and D_B for the moral element M_i as:

$$\textit{m-distance}_{M_i}^{(D_A,D_B)} = d(\mathbf{i}^{(D_A,M_i)}, \mathbf{i}^{(D_B,M_i)})/n, \quad (1)$$

where d is Euclidean distance. The common set of words W offers a common reference point for measuring the distance between lexicons—however, we employ the full domain vocabulary to perform qualitative comparisons between domains (Section 5.4). We normalize the distance by n to reward domains with larger sets of common words. For a domain pair we compute ten *m*-distances, one for each M_i.

Domain Lexicons To compare two domain lexicons, Tomea computes a *d*-distance. The *d*-distance between two domains D_A and D_B is the Euclidean norm of the vector of all *m*-distances computed between the two domains. Intuitively, the Euclidean norm represents the length of the vector of *m*-distances—the larger the *m*-distances between two domains, the larger the *d*-distance. For MFT, with ten moral elements, *d*-distance is:

$$\textit{d-distance}^{(D_A,D_B)} = \sqrt{\sum_{i=1}^{10} (\textit{m-distance}_{M_i}^{(D_A,D_B)})^2} \quad (2)$$

4 Experiment Design

We evaluate Tomea on MFTC (Hoover et al., 2020). Using Tomea, we generate moral and domain lexicons for the seven MFTC domains and perform pairwise comparisons, obtaining 10 *m*-distances and one *d*-distance per comparison. The *m*-distances and *d*-distances are intended to compare the classifiers' representation of moral rhetoric across domains. We perform two types of evaluation to inspect the extent to which these distances capture the differences in moral expression across domains. We also perform a qualitative analysis to find fine-grained differences across domains.

4.1 Dataset

MFTC consists of 35,108 tweets, divided into seven datasets, each corresponding to a different subject: All Lives Matter (**ALM**), Baltimore protests (**BLT**), Black Lives Matter (**BLM**), hate speech and offensive language (**DAV**) (Davidson et al., 2017), 2016 presidential election (**ELE**), MeToo movement (**MT**), and hurricane Sandy (**SND**). Since MFTC consists of datasets from different domains but annotated with the same moral theory, we can perform cross-domain comparisons on the corpus.

Each tweet is labeled with one or more of the 10 moral elements of MFT or a *nonmoral* label. Thus, a tweet can have 11 possible labels. To compensate for the subjectivity of morality annotation, each tweet is annotated by multiple annotators (ranging from 3 to 8). The authors of MFTC apply a majority vote to select the definitive label(s) of each tweet, and tweets with no majority label are labeled as nonmoral. Table 2 shows the distribution of labels and the MeanIR, a measure of label imbalance (Charte et al., 2015) for MFTC. The imbalance is high for some domains, which turns out to be an important factor in the cross-domain comparisons.

4.2 Model Training

We treat morality classification as a multi-class multi-label classification with BERT (Devlin et al., 2019), similar to the recent approaches (Liscio et al., 2022a; Alshomary et al., 2022; Kiesel et al.,

Element	ALM	BLT	BLM	DAV	ELE	MT	SND
Care	456	171	321	9	398	206	992
Harm	735	244	1037	138	588	433	793
Fairness	515	133	522	4	560	391	179
Cheating	505	519	876	62	620	685	459
Loyalty	244	373	523	41	207	322	415
Betrayal	40	621	169	41	128	366	146
Authority	244	17	276	20	169	415	443
Subversion	91	257	303	7	165	874	451
Purity	81	40	108	5	409	173	56
Degradation	122	28	186	67	138	941	91
Nonmoral	1744	3826	1583	4509	2501	1565	1313
Total	4424	5593	5257	5358	4961	4591	4891
MeanIR	11.5	51.3	5.4	344.8	9.6	4.0	6.4

Table 2: Labels distribution per domain of the MFTC.

2022; Huang et al., 2022). We create seven models (one per domain) using the *sequential training* paradigm (Lourie et al., 2021). That is, for each domain, the model is first pre-trained on the other six domains, and then continued training on the seventh. We choose this paradigm since: (1) it is shown to offer the best performance in transfer learning (Lourie et al., 2021; Liscio et al., 2022a), and (2) it represents a realistic scenario, where it is fair to assume that several annotated datasets are available when a novel dataset is collected. Appendix A includes additional details on training.

4.3 Pairwise Comparisons

We employ Tomea to perform pairwise comparisons across the seven domains. First, we generate a moral lexicon for each of the ten moral elements in each of the seven domains (we neglect the *nonmoral* label as it does not expose moral rhetoric). This yields 70 moral lexicons. For each moral element, we perform pairwise comparisons across the seven domains, resulting in 21 *m*-distances per element. Finally, we perform pairwise comparisons of the seven domain lexicons to obtain 21 *d*-distances.

4.4 Evaluation

We evaluate the extent to which *m*-distances and *d*-distances are predictive of differences in moral expression across domains. First, we perform a crowd evaluation to compare moral lexicons and their related *m*-distances. Then, we evaluate domain lexicons and *d*-distances by correlating them to the out-of-domain performances of the models.

4.4.1 Crowd Evaluation

We recruited human annotators on the crowdsourcing platform Prolific[2] to evaluate the comparisons of moral lexicons generated for the same moral element across domains (i.e., the *m*-distances). We designed our annotation task with the covfee annotation tool (Vargas Quiros et al., 2022). The Ethics Committee of the Delft University of Technology approved this study, and we received an informed consent from each subject.

Tomea provides *m*-distances that indicate the distance between domains for each moral element. We evaluate whether humans reach the same conclusions of domain similarity given the moral lexicons generated by Tomea. However, directly providing a distance or similarity between two domains is a challenging task for humans since it lacks a reference point for comparison. Thus, we re-frame the task as a simpler comparative evaluation.

Crowd task We represent each moral lexicon through a word bubble plot, where the 10 most impactful words are depicted inside bubbles scaled by word impact (Figure 2 shows an example). A crowd worker is shown three word bubbles, generated for the same moral element in three domains, D_A, D_B, and D_C. We ask the worker to indicate on a 6-point Likert scale whether D_A is more similar to D_B or D_C based on the shown word bubbles. Appendix B shows a visual example of the task.

Figure 2: Word bubble plot used in the crowd evaluation for the moral element *betrayal* in the BLT domain.

We fix one domain as D_A and choose all possible combinations of the other six domains as D_B and D_C, leading to $(6 * 5)/2 = 15$ combinations. We employ each of the seven domains as D_A, leading to 105 combinations. We generate these combinations for each of the ten moral elements, resulting in 1050 unique tasks. To account for the subjectivity in the annotation, we ensure that each task

[2] www.prolific.co

is performed by three annotators, pushing the total number of required annotations to 3150. Each annotator performed 20 tasks, resulting in a total of 159 annotators. We included four control tasks in each annotator's assignment. Appendix B provides additional details on the crowd study.

Evaluation To compare the results of Tomea and the crowd annotations, we compute the correlation between *m*-distances and crowd answers. Since the Shapiro test showed that the crowd answers are not normally distributed, we choose Spearman correlation in which only the rank order matters.

In the crowd task, workers choose domain similarity on a six-point Likert scale. Given a domain triple (D_A, D_B, D_C), we represent the three choices indicating D_A to be more similar to D_B than D_C as $[-2.5, -1.5, -0.5]$, and D_A to be more similar to D_C than D_B as $[0.5, 1.5, 2.5]$. For each annotation task, we average the answers received by the three annotators that performed it.

In contrast, Tomea computes scores for a domain pair. To compare Tomea's output with the output of the crowd workers, we transform the results of Tomea into the same triples evaluated in the crowd task. To do so, for a domain triple (D_A, D_B, D_C) and a moral element M_i, we compute:

$$\mathcal{S} = \textit{m-distance}_{M_i}^{(D_A, D_B)} - \textit{m-distance}_{M_i}^{(D_A, D_C)}$$

As *m*-distances reflect distance between domains, a negative $\mathcal{S}$ indicates that D_A is more similar to D_B than D_C and a positive $\mathcal{S}$ indicates that D_A is more similar to D_C than D_B. We correlate $\mathcal{S}$ and crowd answers for all 1050 annotated combinations.

4.4.2 Out-of-Domain Performance

The *d*-distances computed by Tomea indicate the similarity between two domains. The more similar the two domains are, the better we expect the out-of-domain performance to be. That is, if domains D_A and D_B are similar, we expect a model trained on D_A to have good classification performance on D_B, and vice versa. Thus, we evaluate the *d*-distances by correlating them to the out-of-domain performances of the models, computed by evaluating each model on the remaining six domains.

5 Results and Discussion

First, we describe the pairwise comparisons resulting from Tomea. Then, we describe the results from the evaluations. Finally, we perform a qualitative analysis to provide fine-grained insights.

5.1 Cross-Domain Comparisons

For each moral element we perform pairwise comparisons across the seven domains, resulting in 21 *m*-distances per element. We aggregate the moral lexicons obtained for the ten moral elements to attain seven domain lexicons. We perform pairwise comparisons across the seven domain lexicons to obtain 21 *d*-distances, which we display in Figure 3 as a 7x7 symmetric matrix. For readability, we show the scores multiplied by 100.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	6.24	4.64	6.84	5.29	5.38	5.55
BLT	6.24	-	6.23	6.09	5.37	5.50	5.56
BLM	4.64	6.23	-	6.27	4.68	5.14	5.25
DAV	6.84	6.09	6.27	-	5.96	6.54	6.80
ELE	5.29	5.37	4.68	5.96	-	4.72	4.62
MT	5.38	5.50	5.14	6.54	4.72	-	4.96
SND	5.55	5.56	5.25	6.80	4.62	4.96	-

Table 3: *d*-distances with moral rhetoric distance between domains. Darker color depicts smaller distance.

First, we observe that the *d*-distances have a small magnitude and variation. This is due to the normalization in Equation 1 (the length of the shared vocabulary, n, is in the order of thousands).

Second, we intuitively expect the moral rhetoric in the domains ALM and BLM to be relatively similar compared to other domain pairs involving ALM or BLM. The *d*-distances support this intuition.

Third, the BLT and DAV domains have the largest overall distances from the other domains. This can be explained by their label distribution (Table 2), which leads to poor accuracy in predicting moral elements (Liscio et al., 2022a; Huang et al., 2022). As these two domains contain fewer tweets labeled with moral elements, the moral lexicons inferred in these domains are of low quality. This may explain why BLM and BLT, both domains involving protests, do not have a low *d*-distance.

Finally, we caution that the *d*-distances in Table 3 are aggregated across moral elements. Although the *d*-distances provide some intuition, the underlying *m*-distances provide more fine-grained information (Section 5.4 and Appendix C).

5.2 Crowd Evaluation

Recall that the crowd evaluation consisted of 1050 domain triples and each triple was annotated by three annotators. The resulting Intra-Class Correlation (ICC) between the annotators, an inter-rater reliability (IRR) metric for ordinal data, was 0.66, which can be considered good but not excellent

(Hallgren, 2012). This shows that crowd workers did not annotate randomly, but can interpret the moral elements differently. Such subjectivity is inevitable when annotating constructs such as morality (Hoover et al., 2020; Liscio et al., 2022b).

We compute the Spearman's rank correlation (ρ) between the crowd annotations and the *m*-distances as described in Section 4.4.1. Table 4 groups the correlations by domains and moral elements. The mean correlation (without any grouping) is 0.4.

Domain	ρ
ALM	0.38
BLT	0.31
BLM	0.43
DAV	0.50
ELE	0.39
MT	0.42
SND	0.31
Average	0.39 ± 0.07

(a) Correlation by domain.

Moral Element	ρ
Care	0.34
Harm	0.57
Fairness	0.74
Cheating	0.23
Loyalty	0.52
Betrayal	0.63
Authority	0.20
Subversion	0.51
Purity	-0.05
Degradation	0.35
Average	0.4 ± 0.24

(b) Correlation by element.

Table 4: Correlation between crowd annotations and *m*-distances, divided by domain and moral element.

We make two observations. First, despite the subjectivity and complexity in comparing moral lexicons, Tomea's results are positively and moderately correlated with human judgment. This shows that Tomea can quantify the differences in how moral elements are represented across domains.

Second, although the agreement between Tomea and humans is consistent across domains, there are large variations across moral elements—spanning strong (e.g., *fairness*), weak (e.g., *authority*), and negligible (e.g., *purity*) correlations. Although the lack of annotations for some moral elements in the corpus has likely influenced these results, such variations cannot be solely explained by the label imbalance. In fact, there is only a weak correlation ($\rho = 0.24$) between the average number of annotations of a moral element across domains (Table 2) and the results in Table 4b. Thus, we conjecture that other factors influence these variations. On the one hand, some moral elements could be more difficult to identify in text than others (Araque et al., 2020; Kennedy et al., 2021). On the other hand, a strong correlation for a moral element could suggest clear differences in representing that element across domains, which both humans and Tomea recognize. Instead, a weak correlation indicates that the agreement between Tomea and humans is almost random, which could suggest that the differences across domains are small or hard to identify.

5.3 Out-of-Domain Performance

To compare the domain lexicons, we compare the *d*-distances to the out-of-domain performance of the models (Section 4.4.2). Table 5 shows the out-of-domain macro F_1-scores of the models. The rows indicate the domain on which the model was trained, and the columns indicate the domain on which the model was evaluated. For each target domain (i.e., each column) we highlight in bold the source domain that performed best.

Target → Source ↓	**ALM**	**BLT**	**BLM**	**DAV**	**ELE**	**MT**	**SND**
ALM	-	48.2	83.7	11.0	68.6	61.9	61.2
BLT	58.5	-	71.6	10.7	56.2	52.2	52.7
BLM	**74.0**	49.9	-	**12.8**	**75.5**	64.3	64.9
DAV	49.3	31.7	64.5	-	37.9	40.4	37.1
ELE	73.9	53.6	87.6	11.9	-	67.0	67.5
MT	71.5	**56.2**	84.4	11.5	72.9	-	**72.3**
SND	73.4	51.6	**88.0**	12.7	72.1	**67.7**	-

Table 5: Macro F_1-scores of models trained on the source domain and evaluated on the target domain.

We notice that no single domain stands out as the best source for all targets. Thus, the choice of the source domain influences a model's out-of-domain performance in a target domain. Hence, we investigate whether the distances Tomea computes are indicative of the out-of-domain performances.

We find a strong negative correlation ($\rho = -0.79$) between the *d*-distances in Table 3 and the out-of-domain F_1-scores in Table 5. Thus, the smaller the *d*-distance between domains, the higher the out-of-domain performance. This demonstrates that Tomea can provide valuable insights on the out-of-domain performance of a model. To scrutinize this result further, we group the correlations by domain in Table 6. There is a moderate to strong negative correlation in all domains except BLT and DAV. We believe that these exceptions are because of the label imbalance and poor model performance in these two domains mentioned in Section 5.1.

	ALM	**BLT**	**BLM**	**DAV**	**ELE**	**MT**	**SND**
ρ	-1.0	0.43	-0.89	0.31	-0.71	-0.83	-0.54

Table 6: Correlation between Tomea results and out-of-domain performance of the models, divided by domain.

5.4 Qualitative Analysis

In addition to quantitative analyses, Tomea enables deep qualitative analyses of the moral expression across domains. In this section, we show examples of (1) words that have high impact on the same moral element across domains, (2) words that have largely different impact on the same moral element across domains, and (3) words that have relatively high impact on two different moral elements in two different domains. Then, we show an example procedure for analyzing the differences between two domains. All lexicon values indicated in these analyses are normalized using the z-score.

First, Tomea can detect words that have a high impact on a moral element across domains. For example, the word 'equality' has high impact on *fairness* in both ALM (21.9) and BLM (27.7) domains; similarly, the word 'fraudulent' has high impact on *cheating* in both domains (22.6 for ALM and 16.0 for BLM). Such consistencies with a large number of words shared between the domains show a consistent moral rhetoric across the domains.

Second, Tomea can detect words whose impact on a moral element largely varies across domains. This information offers a qualitative perspective on the domain dependency of moral elements. For example, ALM and BLM are two of the most similar domains (Table 3). Yet, Tomea indicates that the word 'treason' has a relatively low impact on the moral element of *betrayal* in ALM (2.6) but a considerably higher impact in BLM (24.6); similarly, the word 'brotherhood' has a high impact on *purity* in ALM (26.9) but a comparably lower impact in BLM (8.3). Another interesting comparison can be found between the SND and BLT domains, where the word 'embarrassing' has negligible impact on *degradation* in SND (-0.1) but a high impact in BLT (27.2). These differences can be explained by anecdotal knowledge—that is, the word 'embarrassing' is not relevant for *degradation* in the Hurricane Sandy relief domain, but it is more relevant in the domain of the Baltimore protests.

Third, Tomea can indicate how a word's impact can vary across moral elements, depending on the domain. For example, the word 'crook' has comparable impacts on *cheating* in the ELE domain (3.1) and on *degradation* in the MT domain (3.9); similarly, the word 'looting' has a significant impact on *harm* in ALM (3.5) and on *cheating* in ELE (6.4). These examples demonstrate why domain is crucial in interpreting the moral meaning of a word.

Finally, Tomea facilitates fine-grained comparisons among specific domains of interest. Take ALM and BLM, two very similar domains according to Table 3, for instance. Generally, the *m*-distances of the moral elements are low for these two domains, as shown in Table 7. However, the *m*-distances for *authority* and *subversion* are relatively higher than others. We can inspect this further using the moral lexicons generated by Tomea. For example, in *subversion*, words such as 'overthrow' and 'mayhem' have a high impact in ALM, whereas words such as 'encourage' and 'defiance' have a high impact in BLM. This is in line with our intuition that *subversion* has different connotations in the two domains—whereas subversion is negative in ALM, it is instead encouraged in BLM.

Moral Element	***m*-distance**	**Moral Element**	***m*-distance**
Care	1.62	Harm	1.15
Fairness	1.49	Cheating	1.30
Loyalty	1.54	Betrayal	1.34
Authority	1.80	Subversion	1.85
Purity	1.10	Degradation	1.30

Table 7: The *m*-distances between ALM and BLM.

The analyses above are not meant to be exhaustive. We pick examples of moral elements, domains, and words to demonstrate the fine-grained analyses Tomea can facilitate. Our observations, considering that we only analyzed a few examples, may not be significant in themselves. Further, these observations may change with more (or other) data.

6 Conclusions and Directions

Tomea is a novel method for comparing a text classifier's representation of morality across domains. Tomea offers quantitative measures of similarity in moral rhetoric across moral elements and domains. Further, being an interpretable method, Tomea supports a fine-grained exploration of moral lexicons. Tomea is generalizable over a variety of classification models, domains, and moral constructs.

The similarities computed by Tomea positively correlate with human annotations as well as the out-of-domain performance of morality prediction models. Importantly, Tomea can shed light on how domain-specific language conveys morality, e.g., the word 'brotherhood' has a high impact on moral elements in the ALM domain, whereas the word 'treason' has a high impact in the BLM domain.

Tomea can be a valuable tool for researchers and

practitioners. It can be used to study how a text classifier represents moral rhetoric across personal, situational, and temporal dimensions, and across different types of moral values (Pommeranz et al., 2012; Liscio et al., 2022b). Tomea can support societal applications such as modeling stakeholders' preferences on societal issues (Mouter et al., 2021; Siebert et al., 2022; Liscio et al., 2023), analyzing the impact of events like the COVID-19 pandemic (van de Poel et al., 2022), and predicting violent protests (Mooijman et al., 2018). Finally, Tomea can assist NLP researchers in generating morally aligned text (Ammanabrolu et al., 2022; Bakker et al., 2022) that is domain specific.

A key direction to improve Tomea is incorporating refined explanations, e.g., by rule-based inferences (Zhou et al., 2022). Additional distance metrics and normalization procedures may also provide a more accurate lexicon comparison. Finally, the qualitative analysis that we performed could be systematized as a methodology for analysts.

7 Ethical Considerations and Limitations

There is a growing interest in investigating human morality in text (Russell et al., 2015; Gabriel, 2020). However, like most technologies, morality classification can be misused, especially targeting sensitive features including ethnicity and political orientation (Kalimeri et al., 2019a; Talat et al., 2022). For instance, authorities in non-liberal countries could use Tomea to identify repressed minorities by detecting moral language that diverges from the expected moral rhetoric. Ongoing research is investigating such issues, e.g., by creating methods that mitigate bias and unfairness by design (Dinan et al., 2020; Vargas and Cotterell, 2020).

We discuss three main limitations of our analyses related to the corpus we use (MFTC). First, MFTC is composed of English tweets, and we employ a version of BERT that was pre-trained on large-scale English data. Our experiments show that Tomea produces insightful results under these conditions. However, the performance of Tomea with models pre-trained on smaller datasets, e.g., datasets for morphologically richer languages, remains to be investigated. Further, the scalability of Tomea to longer text formats (e.g., news articles) and different mediums of communication (e.g., surveys) is yet to be explored.

Second, the tweets in the MFTC were collected using the Twitter API, which only yields public posts. Thus, following Twitter's Terms of Service, deleted content will not be available (limiting the reproducibility of any Twitter-based study). Further, the demographic and cultural distribution of Twitter users may not be representative of the general population, In addition, we required the crowd workers involved in the evaluation to be fluent in English, and their demographic distribution (Appendix B.3) is skewed towards Europe. These factors could possibly lead to the perpetuation of Western values and biases (Mehrabi et al., 2021) in our analyses. Additional experiments are needed to investigate whether Tomea would produce insightful results when applied on a dataset collected on a more extensive slice of the population, with a broader set of linguistical expressions.

Third, the MFTC is focused on US-centric topics. However, when recruiting annotators for our crowd evaluation, we did not require familiarity with such topics. Even though the annotators were not exposed to the original tweets but to a processed version of the dataset (i.e., the output of Tomea, see Section 4.4.1), the potential lack of familiarity may have influenced the evaluation results.

Finally, we remind that Tomea's d-distances measure how (dis-)similar two domains are, and are thus not a (binary) judgment of (dis-)similarity. Further, two corpora collected in the same domain (e.g., two datasets on BLM protests) will likely not have a d-distance of 0. It is left to the user to judge the similarity of the two corpora, supported by Tomea's quantitative and qualitative metrics.

Acknowledgments

This research was partially supported by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organization for Scientific Research. Oscar Araque acknowledges the funding by the European Union's Horizon 2020 research and innovation program under grant agreement 962547 (PARTICIPATION).

References

Tareq Al-Moslmi, Nazlia Omar, Salwani Abdullah, and Mohammed Albared. 2017. Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review. *IEEE Access*, 5:16173–16192.

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*, ACL '22, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to Social Norms and Values in Interactive Narratives. In *Proceedings of the 2022 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 5994–6017, Seattle, USA. Association for Computational Linguistics.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:1–11.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. LibertyMFD: A Lexicon to Assess the Moral Foundation of Liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 154–160, New York, NY, USA. Association for Computing Machinery.

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: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, DeeLIO '22, pages 33–41, Dublin, Ireland and Online. Association for Computational Linguistics.

Mohamed Bahgat, Steven R. Wilson, and Walid Magdy. 2020. Towards Using Word Embedding Vector Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, pages 919–923, Atlanta, Georgia. AAAI Press.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, NeurIPS '22, pages 38176–38189. Curran Associates, Inc.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Johan Brännmark. 2015. Moral disunitarianism. *The Philosophical Quarterly*, 66(264):481–499.

Jay Carriere, Hareem Shafi, Katelyn Brehon, Kiran Pohar Manhas, Katie Churchill, Chester Ho, and Mahdi Tavakoli. 2021. Case Report: Utilizing AI and NLP to Assist with Healthcare and Rehabilitation During the COVID-19 Pandemic. *Frontiers in Artificial Intelligence*, 4(2):1–7.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, AACL '20, page 447–459, Suzhou, China.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '19, page 4171–4186.

Janis L Dickinson, Poppy McLeod, Robert Bloomfield, and Shorna Allred. 2016. Which moral foundations predict willingness to make lifestyle changes to avert climate change in the USA? *PLoS ONE*, 11(10):1–11.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 314–331.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 653–670, Online. Association for Computational Linguistics.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, pages 3730–3736.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

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, Amsterdam, the Netherlands.

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.

Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*, 8(1):23–34.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 595–605, Austin, Texas, USA.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In *Proceedings of the 2021 International Conference on Learning Representations*, ICLR '21, pages 1–29.

Patrick L. Hill and Daniel K. Lapsley. 2009. Persons and situations in the moral domain. *Journal of Research in Personality*, 43(2):245–246.

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.

Xiaolei Huang, Alexandra Wormley, and Adam Cohen. 2022. Learning to Adapt Domain Shifts of Moral Values via Instance Weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, pages 121–131. Association for Computing Machinery.

Kyriaki Kalimeri, Mariano G. Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019a. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445.

Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019b. Human values and attitudes towards vaccination in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 248–254.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. Moral Concerns are Differentially Observable in Language. *Cognition*, 212:104696.

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*, ACL '22, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Ilir Kola, Ralvi Isufaj, and Catholijn M. Jonker. 2022. Does Personalization Help? Predicting How Social Situations Affect Personal Values. In *HHAI2022: Augmenting Human Intellect*, pages 157–169.

Alex Gwo Jen Lan and Ivandré Paraboni. 2022. Text- and author-dependent moral foundations classification. *New Review of Hypermedia and Multimedia*, 0(0):1–21.

Enrico Liscio, Alin E. Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022a. Cross-Domain Classification of Moral Values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '22, pages 2727–2745, Seattle, USA. Association for Computational Linguistics.

Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. 2023. Value inference in sociotechnical systems: Blue sky ideas track. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1–7, London, United Kingdom. IFAAMAS.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022b. What Values Should an Agent Align With? *Autonomous Agents and Multi-Agent Systems*, 36(23):32.

Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. 2018. On interpretation of network embedding via taxonomy induction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pages 1812–1820. ACM.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13480–13488.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *booktitle = Advances in Neural Information Processing Systems,*, NeurIPS '17, pages 1208–1217, Long Beach, CA, USA.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6).

Omid Mohamad Beigi and Mohammad H. Moattar. 2021. Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. *Knowledge-Based Systems*, 213:106423.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 2(6):389–396.

Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. 2021. Public Participation in Crisis Policymaking. How 30,000 Dutch Citizens Advised Their Government on Relaxing COVID-19 Lockdown Measures. *PLoS ONE*, 16(5):1–42.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL '21, pages 5356–5371, Online. Association for Computational Linguistics.

Matheus C. Pavan, Vitor G. Santos, Alex G. J. Lan, Joao Martins, Wesley Ramos Santos, Caio Deutsch, Pablo B. Costa, Fernando C. Hsieh, and Ivandre Paraboni. 2020. Morality Classification in Natural Language Text. *IEEE Transactions on Affective Computing*, 3045(c):1–8.

Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M. Jonker. 2012. Elicitation of Situated Values: Need for Tools to Help Stakeholders and Designers to Reflect and Communicate. *Ethics and Information Technology*, 14(4):285–303.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text. *European Journal of Personality*, 34(5):885–902.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wager. 2018. Deconfounded Lexicon Induction for Interpretable Social Science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '18, pages 1615–1625, New Orleans, Louisiana, USA.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. ValueNet: A New Dataset for Human Value Driven Dialogue System. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, AAAI '22, pages 11183–11191.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144.

Stuart J. Russell, Daniel Dewey, and Max Tegmark. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4):105–114.

Chelsea Schein. 2020. The Importance of Context in Moral Judgments. *Perspectives on Psychological Science*, 15(2):207–215.

Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture*, 2(1):1–20.

Luciano C. Siebert, Enrico Liscio, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon L. Spruit, Jeroen van den Hoven, and Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *HHAI2022: Augmenting Human Intellect*, pages 114–127, Amsterdam, the Netherlands. IOS Press.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 769–779, Seattle, USA.

Mario Triola. 2017. *Elementary Statistics*, 13th edition. Pearsons.

Ibo van de Poel, Tristan de Wildt, and Dyami van Kooten Pássaro. 2022. COVID-19 and Changing Values. In *Values for a Post-Pandemic Future*, pages 23–58. Springer International Publishing.

Francisco Vargas and Ryan Cotterell. 2020. Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 2902–2913.

Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. 2022. Covfee: an extensible web framework for

continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, pages 265–293. PMLR.

Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C. Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. 2019. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *npj Digital Medicine*, 2(130):1–7.

Garrett Wilson and Diane J. Cook. 2020. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5).

Steven R. Wilson, Yiting Shen, and Rada Mihalcea. 2018. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '18, pages 455–470, St. Petersburg, Russia. Springer.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 301–310, Berlin, Germany. Association for Computational Linguistics.

Yilun Zhou, Marco Tulio Ribeiro, and Julie Shah. 2022. Exsum: From local explanations to model understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22, pages 5359–5378, Seattle, USA. Association for Computational Linguistics.

A Experimental Details

We provide here all the information needed for reproducing our experimental results. Code and the complete set of results are provided as supplemental material. The models cannot be shared due to upload size limit, thus will be shared at publication.

A.1 Data Preprocessing

We preprocess the tweets by removing URLs, emails, usernames and mentions. Next, we employ the Ekphrasis package[3] to correct common spelling mistakes and unpack contractions. Finally, emojis are transformed into their respective words using the Python Emoji package[4].

A.2 Hyperparameters

To select the hyperparameters, we trained and evaluated the model on the entire MFTC corpus with 10-fold cross-validation. Table A1 shows the hyperparameters that were compared in this setting, highlighting in bold the best performing option that we then used in the experiments described in the paper. If a parameter is not present in the table, the default value supplied by the framework was used.

Hyperparameters	Options
Model name	**bert-base-uncased**
Number of parameters	**110M**
Max sequence length	**64**
Epochs	2, **3**, 5
Batch size	**16**, 32, 64
Dropout	0.05, **0.1**, 0.02
Optimizer	**AdamW**
Learning Rate	**5∗10^{-5}**
Loss function	**Binary Cross Entropy**

Table A1: Hyperparameters tested and selected.

A.3 Model Training

As introduced in Section 4.2, we trained seven models on the seven domains of the MFTC, respectively. Each model was first trained on the remaining six domains, and then continued training on the domain under analysis. The training on the seventh domain was performed on 90% of the domain, leaving 10% out for evaluation. Table A2 shows the performances of the models on the domains portions left out for evaluation.

[3] https://github.com/cbaziotis/ekphrasis

[4] https://pypi.org/project/emoji/

	ALM	BLT	BLM	DAV	ELE	MT	SND
F_1-score	70.3	32.1	85.3	8.7	64.8	62.3	53.9

Table A2: Models performance (macro F_1-score).

A.4 Computing Infrastructure

The following are the main libraries and computing environment used in our experiments.

- PyTorch: 1.8.1
- Hugginface's Transformers: 4.6.0
- NVIDIA GeForce RTX 2080 Ti GPU
- CUDA: 11.2
- cuDNN: 8.1.1.33
- SHAP: 0.40.0

We spent 7 GPU hours to train the seven models used in the experiments. We spent 70 CPU hours to generate the moral lexicons.

A.5 Random Seeds

In our experiments, to control for randomness, we fixed the random seeds in the following libraries:

- Python (`random.seed`)
- NumPy (`numpy.random.seed`)
- PyTorch (`torch.manual_seed`)
- CUDA (`torch.cuda.manual_seed_all`)

A.6 Artifacts Usage

We have mainly used three artifacts in this research: the MFTC (Hoover et al., 2020), SHAP (Lundberg and Lee, 2017), and BERT (Devlin et al., 2019).

The MFTC was collected with the intent of facilitating NLP research on morality. It can be downloaded[5] and used under the Creative Commons Attribution 4.0 license.

SHAP was intended to explain the output of any machine learning model. Thus, we are using it as originally intended, under its MIT license[6].

BERT was created with the intent of performing, among others, text classification. Thus, we are using it as originally intended, under its Apache 2.0 distribution license[7].

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

[6] https://github.com/slundberg/shap/blob/master/LICENSE

[7] https://github.com/google-research/bert/blob/master/LICENSE

B Crowd Evaluation

Section 4.4.1 introduces the crowd experiment. We first opened a pilot annotation job on Prolific for nine users with an expected completion time of 25 minutes. The average completion time was 21 minutes and the average ICC 0.61. These results encouraged us to proceed with the rest of the experiment. Ultimately, the average time spent by a crowd worker on a job was 22 minutes ($\pm$ 12 minutes SD). Each worker was paid £3.75 (at the rate of £9/h as per Prolific suggestion of fair retribution).

B.1 Annotation Job Layout

Upon taking the annotation job on Prolific, workers were redirected to a web application hosted on our servers. Here, after accepting the informed consent form, they were asked demographic questions and then were given a brief introduction to the annotation tasks and the moral elements involved. Informed consent form, instructions, and all word bubbles are provided as supplemental material.

Figure B2 shows an example of an annotation task. In each individual task, annotators needed to indicate whether the word bubble describing domain D_A was more similar to the one describing domain D_B or D_C. The annotators were given the following six options on a Likert scale:

1. A is clearly more similar to B (than to C)
2. A is more similar to B (than to C)
3. A is slightly more similar to B (than to C)
4. A is slightly more similar to C (than to B)
5. A is more similar to C (than to B)
6. A is clearly more similar to C (than to B)

After the initial instructions, each annotator was guided through four sections. Each section contained five tasks where all word bubbles were generated for the same moral element (but multiple different domains), plus one control task (as described in Section B.2). Before each section, the annotator was introduced to the moral element concerned in the following section. Thus, each annotator was introduced to four different moral elements. These elements were chosen from two different moral foundations, for a total of two moral foundations per annotator. For instance, one annotation job could be composed of four annotation sections corresponding to the moral elements of *care*, *harm*, *authority*, and *subversion*, resulting in 24 annotations tasks (including four control tasks).

B.2 Quality Control

The crowd workers were required to be fluent in English and have submitted at least 100 Prolific jobs with at least 95% acceptance rate. We included four control tasks, one per section. In each, the word bubbles describing D_A and D_B were identical, and different from the word bubble describing D_C.

A total of 186 workers completed the job. Using the Likert options enumeration introduced in Section B.1, we included a worker's job in our analysis only if (1) all four control tasks were answered with options 1, 2, or 3; and (2) at least two control tasks were answered with options 1 or 2. These criteria were set before any analysis of crowd work was done. Of the 186 workers, 159 satisfied the criteria above.

B.3 User demographics

Upon giving informed consent, workers were asked the following demographic information:

- What is your age?
- What gender do you identify as?
- Where is your home located?
- What is the highest degree or level or education you have completed?

Figure B1 shows the demographics of the 159 users whose submissions were considered in the study.

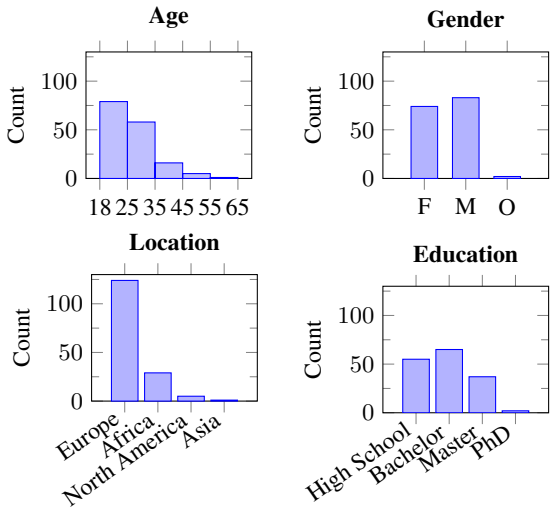

Figure B1: Demographics of crowd workers.

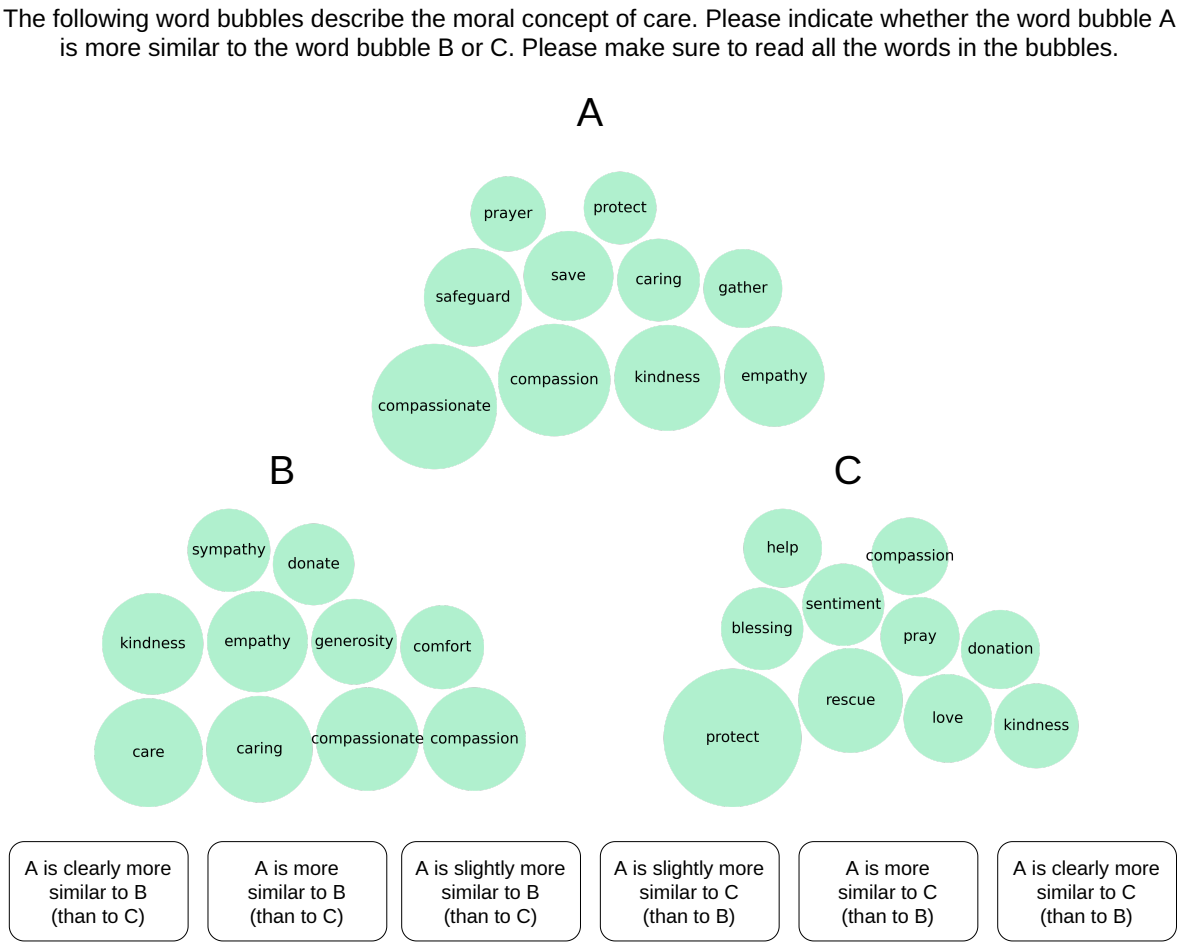

Figure B2: The annotator is asked to take a choice on a 6-points Likert scale based on the shown word bubbles.

C Extended Results

C.1 *m*-distances

In Table 3 we show the *d*-distances describing the distance between domains. In tables C1a to C1j we display the *m*-distances describing the distance between domains for each moral element. For readability, we show the scores multiplied by 100.

The most apparent consideration is that moral expression similarity is not consistent across domains, but rather depends on the moral element under analysis. In Section 5.4 we provide examples on how to explore such fine-grained differences across domains. On top of the explored cases, another insightful example is represented by two domains that ranked with a higher distance, ALM and SND. Nevertheless, the domains ranked relatively more similar in the *care* element. Let us inspect closely the moral lexicons generated for *care* for ALM and SND. At first, we notice some differences, such as the words 'rescue' and 'donation' that are specific to the SND domain, being especially relevant in a hurricane relief domain. However, we also notice many similarities, such as the words 'protect' and 'compassion', typical for describing in-group care.

C.2 Correlation by Domain and Element

Table C2 shows the Spearman correlation (ρ) by moral element and domain. We notice that ρ is generally consistent across moral elements—for instance, the elements of *fairness* and *betrayal* have the highest ρ, while *purity* have the lowest. However, there are some exceptions. SND has a comparatively low ρ for *harm*, and MT for *subversion*, despite having a large number of annotations (Table 2). A possible reason is that the expression of these elements in these domains is less domain specific than in other domains, leading to lower ρ with crowd intuition. Instead, DAV has a high ρ for *harm* and *betrayal*. This can be explained by the nature of the domain (hate speech), which would lead to highly specific lexicons for these elements.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.66	1.62	2.28	1.72	1.51	1.43
BLT	1.66	-	1.68	1.13	1.70	1.62	1.53
BLM	1.62	1.68	-	1.28	1.41	1.98	1.80
DAV	2.28	1.13	1.28	-	1.67	1.96	2.26
ELE	1.72	1.70	1.41	1.67	-	1.82	1.64
MT	1.51	1.62	1.98	1.96	1.82	-	1.61
SND	1.43	1.53	1.80	2.26	1.64	1.61	-

(a) *m*-distances for the *care* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.45	1.15	2.48	1.26	1.23	1.12
BLT	1.45	-	1.44	1.85	1.34	1.33	1.38
BLM	1.15	1.44	-	2.19	1.17	1.14	1.06
DAV	2.48	1.85	2.19	-	1.69	2.15	2.11
ELE	1.26	1.34	1.17	1.69	-	1.11	1.11
MT	1.23	1.33	1.14	2.15	1.11	-	1.02
SND	1.12	1.38	1.06	2.11	1.11	1.02	-

(b) *m*-distances for the *harm* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.17	1.49	2.21	1.65	1.66	1.86
BLT	2.17	-	2.34	2.24	1.96	1.98	2.09
BLM	1.49	2.34	-	2.22	1.67	1.82	1.93
DAV	2.21	2.24	2.22	-	2.14	2.17	2.49
ELE	1.65	1.96	1.67	2.14	-	1.58	1.66
MT	1.66	1.98	1.82	2.17	1.58	-	1.73
SND	1.86	2.09	1.93	2.49	1.66	1.73	-

(c) *m*-distances for the *fairness* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.82	1.30	2.06	1.34	1.60	1.62
BLT	1.82	-	1.84	1.79	1.63	1.62	1.75
BLM	1.30	1.84	-	2.09	1.24	1.35	1.44
DAV	2.06	1.79	2.09	-	2.06	1.98	2.31
ELE	1.34	1.63	1.24	2.06	-	1.23	1.35
MT	1.60	1.62	1.35	1.98	1.23	-	1.47
SND	1.62	1.75	1.44	2.31	1.35	1.47	-

(d) *m*-distances for the *cheating* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.58	1.54	2.46	1.93	2.01	1.96
BLT	1.58	-	1.82	1.36	1.65	1.91	1.73
BLM	1.54	1.82	-	2.35	1.60	1.55	1.99
DAV	2.46	1.36	2.35	-	2.40	2.40	2.75
ELE	1.93	1.65	1.60	2.40	-	1.30	1.68
MT	2.01	1.91	1.55	2.40	1.30	-	1.59
SND	1.96	1.73	1.99	2.75	1.68	1.59	-

(e) *m*-distances for the *loyalty* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.02	1.34	1.75	1.19	1.21	1.13
BLT	2.02	-	1.92	2.04	1.56	1.84	1.73
BLM	1.34	1.92	-	1.69	0.85	1.12	0.90
DAV	1.75	2.04	1.69	-	1.56	1.73	1.61
ELE	1.19	1.56	0.85	1.56	-	1.05	0.87
MT	1.21	1.84	1.12	1.73	1.05	-	0.88
SND	1.13	1.73	0.90	1.61	0.87	0.88	-

(f) *m*-distances for the *betrayal* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.18	1.80	2.21	2.02	1.87	2.00
BLT	2.18	-	2.20	2.31	1.67	1.75	1.65
BLM	1.80	2.20	-	1.81	1.80	1.62	1.79
DAV	2.21	2.31	1.81	-	1.61	2.06	1.82
ELE	2.02	1.67	1.80	1.61	-	1.77	1.63
MT	1.87	1.75	1.62	2.06	1.77	-	1.58
SND	2.00	1.65	1.79	1.82	1.63	1.58	-

(g) *m*-distances for the *authority* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.10	1.85	2.48	1.84	2.17	2.30
BLT	2.10	-	1.98	2.12	1.87	1.78	1.66
BLM	1.85	1.98	-	2.30	1.61	2.05	2.05
DAV	2.48	2.12	2.30	-	2.11	2.00	2.35
ELE	1.84	1.87	1.61	2.11	-	1.72	1.63
MT	2.17	1.78	2.05	2.00	1.72	-	1.84
SND	2.30	1.66	2.05	2.35	1.63	1.84	-

(h) *m*-distances for the *subversion* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	2.86	1.10	1.85	2.14	1.56	2.44
BLT	2.86	-	2.78	2.29	2.24	1.98	2.40
BLM	1.10	2.78	-	1.75	1.79	1.72	1.94
DAV	1.85	2.29	1.75	-	1.61	1.71	2.00
ELE	2.14	2.24	1.79	1.61	-	1.51	1.67
MT	1.56	1.98	1.72	1.71	1.51	-	1.87
SND	2.44	2.40	1.94	2.00	1.67	1.87	-

(i) *m*-distances for the *purity* element.

	ALM	BLT	BLM	DAV	ELE	MT	SND
ALM	-	1.44	1.30	1.65	1.34	1.94	1.03
BLT	1.44	-	1.27	1.77	1.11	1.47	1.40
BLM	1.30	1.27	-	1.89	1.38	1.61	1.21
DAV	1.65	1.77	1.89	-	1.77	2.40	1.44
ELE	1.34	1.11	1.38	1.77	-	1.60	1.09
MT	1.94	1.47	1.61	2.40	1.60	-	1.76
SND	1.03	1.40	1.21	1.44	1.09	1.76	-

(j) *m*-distances for the *degradation* element.

Table C1: *m*-distances for the ten moral elements. Darker color indicates smaller distance between domains.

	Care	Harm	Fairness	Cheating	Loyalty	Betrayal	Authority	Subversion	Purity	Degradation
ALM	0.49	0.53	0.65	0.34	0.49	0.63	0.11	0.47	0.03	0.25
BLT	0.10	0.46	0.73	0.15	0.17	0.59	0.38	0.37	-0.01	0.29
BLM	0.20	0.54	0.66	0.27	0.60	0.67	0.27	0.61	0.16	0.36
DAV	0.43	0.84	0.80	0.18	0.63	0.75	0.39	0.65	-0.26	0.45
ELE	0.41	0.58	0.69	0.43	0.48	0.55	-0.11	0.70	-0.19	0.42
MT	0.36	0.50	0.76	0.24	0.51	0.53	0.25	0.30	0.08	0.44
SND	0.37	0.25	0.73	0.05	0.58	0.69	-0.01	0.47	-0.13	0.21

Table C2: Spearman correlation (ρ) between *m*-distances and crowd results, divided by domain and moral element. Darker color indicates higher correlation.

C.3 Qualitative Analysis

In Section 5.4 we suggest methods for qualitatively comparing moral rhetoric across domains. In particular, we show similarities and differences between two domains, ALM and BLM. These are among the most similar domains for the moral elements of *fairness* (Table C1c) and *cheating* (Table C1d). For both domains, the words 'equality' and 'fraud' are among the most impactful words for the two elements, respectively. In Table C3 we show examples of tweets where these words are used, in order to provide additional context on their usage.

Tweet	Domain	Label
Equality is key. #AllLivesMatter pray over everyone. Cherish your life cause today you never know	ALM	fairness
Praying for Justice and *equality*	BLM	fairness
Of course #AllLivesMatter Shep, you self righteous, dangerously politically correct *fraud* posing as a fair journalist.	ALM	cheating
Shaun King is/was a *fraud* and a liar and deserved to be outed as such. #BlackLivesMatter deserves better.	BLM	cheating

Table C3: Examples of tweets with similar moral rhetoric in the ALM and BLM domains.

On the other hand, ALM and BLM differ in the moral element of *subversion* (Table C1h). Here, words such as 'overthrow' and 'mayhem' have high impact in ALM, whereas words such as 'encourage' and 'defiance' have high impact in BLM. In Table C4 we show examples of tweets where these words are used, in order to provide additional context on their usage.

Tweet	Domain	Label
I am a proponent of civil disobedience and logic driven protest only; not non irrational violence, pillage & *mayhem*!	ALM	subversion
For those who try to confuse acts of *defiance* with deliberate acts of racist terrorism, we pray	BLM	subversion

Table C4: Examples of tweets with different moral rhetoric in the ALM and BLM domains.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 and Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Sections 2, 3, 4, Appendix A

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A6

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A6

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data was collected by Hoover et al. (2020), see Section 4.1. In their paper they discuss the anonimization and filtering process. We further process the tweets by removing URLs, emails, usernames and mentions, as described in Appendix A1.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Details of the artifacts we use are provided by the original authors of MFTC (Hoover et al., 2020) and BERT (Devlin et al., 2019).

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.2 and Appendix A3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4 and Appendix A

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A2 and A4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.2 and Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4,5, Appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B1 and supplemental material (data)

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.4.1 and Appendix B

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B and supplemental material (data)

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 4.4.1 and supplemental material (data)

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B3