Cross-Domain Classification of Moral Values

Enrico Liscio, Alin E. Dondera, Andrei Geadău,
Catholijn M. Jonker, and **Pradeep K. Murukannaiah**
Delft University of Technology, the Netherlands
{E.Liscio,C.M.Jonker,P.K.Murukannaiah}@tudelft.nl
{A.E.Dondera,A.Geadau}@student.tudelft.nl

Abstract

Moral values influence how we interpret and act upon the information we receive. Identifying human moral values is essential for artificially intelligent agents to co-exist with humans. Recent progress in natural language processing allows the identification of moral values in textual discourse. However, domain-specific moral rhetoric poses challenges for transferring knowledge from one domain to another.

We provide the first extensive investigation on the effects of cross-domain classification of moral values from text. We compare a state-of-the-art deep learning model (BERT) in seven domains and four cross-domain settings. We show that a value classifier can generalize and transfer knowledge to novel domains, but it can introduce catastrophic forgetting. We also highlight the typical classification errors in cross-domain value classification and compare the model predictions to the annotators agreement. Our results provide insights to computer and social scientists that seek to identify moral rhetoric specific to a domain of discourse.

1 Introduction

Morality helps humans discern right from wrong. Pluralist moral philosophers argue that human morality can be represented, understood, and explained by a finite number of irreducible basic elements, referred to as *moral values* (Graham et al., 2013). The difference in our preferences over moral values explains how and why we think differently. For instance, both conservatives and liberals may agree that individual welfare is important. However, a conservative, who cherishes the values of freedom and independence, may believe that taxes should be decreased to attain more individual welfare. In contrast, a liberal, who cherishes the values of community and care, may believe that taxes should be increased to obtain welfare (Graham et al., 2009).

It is crucial to understand human morality to develop beneficial AI (Russell et al., 2015; Soares and Fallenstein, 2017). To operate among humans, artificial agents must be able to comprehend and recognize the moral values that drive the differences in human behavior (Akata et al., 2020; Gabriel, 2020). The ability to understand moral rhetoric can be instrumental for, e.g., facilitating human-agent trust (Chhogyal et al., 2019; Mehrotra et al., 2021) and engineering value-aligned socio-technical systems (Ajmeri et al., 2020; Murukannaiah et al., 2020; Serramia et al., 2021; Montes and Sierra, 2021).

There are survey instruments to estimate individual value profiles (Schwartz, 2012; Graham et al., 2013). However, reasoning about moral values is challenging for humans (Le Dantec et al., 2009; Pommeranz et al., 2012). Further, in practical applications, e.g., to conduct meaningful conversations (Tigunova et al., 2019) or to identify online trends (Mooijman et al., 2018), artificial agents should be able to understand moral rhetoric on the fly.

The growing capabilities of natural language processing (NLP) enable the estimation of moral rhetoric from textual discourse (Hoover et al., 2020; Araque et al., 2020; Alshomary et al., 2022; Kiesel et al., 2022). Specifically, a value classifier can be used to identify the moral values underlying a piece of text on the fly. For instance, Mooijman et al. (2018) show that detecting moral values from tweets can predict violent protests.

Existing value classifiers are evaluated on a specific dataset, without re-training or testing the classifier on a different dataset. This shows the ability of the classifier to predict values from text, but not the ability to transfer the learned knowledge across datasets. A critical aspect of moral values is that they are intrinsically linked to the domain under discussion (Pommeranz et al., 2012; Liscio et al., 2021, 2022). Moral value expressions may take different forms in different domains. For example, in the driving domain, the value of safety concerns speed limits and seat belts, but in the COVID-19 domain, safety concerns social distancing and face masks. Further, a word (broadly, language) may trigger different moral rhetoric in different domains. For example, in a libertarian blog, the word 'taxes' may be linked to the authority value, but in a socialist blog it may be linked to the community value. Thus, it is crucial for a value classifier to recognize domain-specific connotations of moral rhetoric.

Collecting and annotating a sufficient amount of training examples in each domain is expensive and time consuming. To reduce the need for new annotated examples, we can pretrain classifiers with similar available annotated data and transfer the acquired knowledge to a novel task—a practice known as *transfer learning* (Ruder, 2019). Despite the benefits, transfer learning poses well-known challenges, including: (1) *generalizability*: how well does a classifier perform on novel data? (2) *transferability*: how well is knowledge transferred from one domain to another? and (3) *catastrophic forgetting*: to what extent is knowledge of a previous domain lost after training in a new domain? These challenges are crucial for value classification because of its domain-specific nature.

We perform the first comprehensive cross-domain evaluation of a value classifier. We employ the Moral Foundation Twitter Corpus (Hoover et al., 2020), consisting of seven datasets spanning different socio-political areas, annotated with the value taxonomy of the Moral Foundation Theory (Graham et al., 2013). Treating each dataset as a domain, we train a deep learning model, BERT (Devlin et al., 2019), in four training settings to evaluate the value classifier's generalizability, transferability, and catastrophic forgetting.

Our experiments show that (1) a value classifier can generalize to novel domains, especially when trained on a variety of domains; (2) initializing a classifier with examples from different domains improves performance in novel domains even when little training data is available in the novel domains; (3) catastrophic forgetting occurs even when training on a small portion of data from the novel domain, and its impact must be considered when training on a novel domain; and (4) in the large majority of cases, in all considered training settings, at least one annotator agrees with the model predictions.

Our investigation is significant because moral rhetoric is seldom explicit in language, but often lies in subtle domain-dependent cues. Understanding whether a classifier can recognize and transfer such hidden patterns across domains is instrumental for the practical use. By unveiling the successes and mistakes of value classifiers in cross-domain settings, we hope to inspire researchers and practitioners to employ value classification responsibly.

2 Background and Data

We introduce the Moral Foundation Theory (MFT) (Graham et al., 2013) and the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020) used in our experiments.

The MFT is a well-established theory of moral values developed by social and cultural psychologists. It argues that human morality is composed of a finite set of innate moral foundations, similar to how the five taste receptors (for sweet, sour, salt, bitter, and umami) combine to yield the tastes we experience. The MFT includes five foundations, each composed of a vice–virtue duality, resulting in the 10 moral values shown in Table 1.

Table 1: The five moral foundations in the MFT

Foundation	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

The MFTC is composed of 35,108 tweets, divided into seven datasets, each corresponding to a topic: 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**). These datasets from complex and diverse socio-political issues allow us to evaluate the transferability by treating each dataset as belonging to a domain.

The tweets were annotated by multiple annotators with the MFT taxonomy. Hoover et al. (2020) provide additional details on the annotation process. They recognize that the vice and the virtue constituting one moral foundation are expressed differently in natural language. For example, an utterance describing a care concern (e.g., taking care of one's offspring) does not necessarily also contain harm expressions. For this reason, each tweet was annotated with all 10 individual moral values plus an additional *nonmoral* label, resulting in 11 possible labels per tweet. Due to the subjective nature of moral values, different annotators may label the same tweet differently. For this reason, Hoover et al. (2020) apply a majority vote to select the definitive label(s) of each tweet. Tweets with no majority label are labeled as nonmoral. Table 2 shows three examples of annotated tweets.

Table 2: Examples of labeled tweets in MFTC

Tweet	**Dataset**	**Labels**
Police lives matter, all lives matter, peace and love people	ALM	care
Which oppression is worse, sexism or racism?	BLM	harm, cheating
Baltimore Police will deliver an update on the #FreddieGray investigation. Listen live on WBAL	BLT	nonmoral

Table 3 shows the distribution of labels. The MeanIR is a measure of imbalance in a dataset (Charte et al., 2015). MeanIR is the mean of IR_l for each label l, where IR_l is the ratio of the number of instances having the majority (i.e., nonmoral) label and the number of instances having label l. The degree of imbalance varies largely across datasets, which is realistic since different domains are likely to have different distributions of moral content.

Table 3: Distribution of labels per dataset of the MFTC

Foundation	**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

3 Experimental Setup

Predicting moral values is a multi-label classification problem. Given a set of textual documents, $\mathcal{T}$, and a set of moral value labels, $\mathcal{L} = (l_1, l_2, \ldots, l_n)$, we wish to learn a mapping $\mathcal{C} : \mathcal{T} \mapsto \mathcal{P}(\mathcal{L})$. Each element in $\mathcal{P}(\mathcal{L})$ is a binary vector, $y = (y_1, y_2, \ldots, y_n)$, where $y_i = 1$ if the corresponding text is labeled with l_i. The mapping $\mathcal{C}$ is learned via BERT (Devlin et al., 2019), a language representation model based on the Transformer architecture (Vaswani et al., 2017). We choose BERT as it represents the state-of-the-art for several NLP tasks, including value classification (Kobbe et al., 2020; Alshomary et al., 2022; Kiesel et al., 2022). We provide additional details, including hyperparameters, in the Appendix. The code is available on GitHub[1].

3.1 Cross-Domain Evaluation

To perform cross-domain evaluation, we partition the MFTC datasets into $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$. We treat $\mathcal{T}_{source}$ as available data and $\mathcal{T}_{target}$ as an incoming dataset from a novel domain. In our experiments, $\mathcal{T}_{target}$ is always composed of one MFTC dataset. We experiment with $\mathcal{T}_{source}$ composed of one, three, and six datasets. We present the results for the setting with six datasets as $\mathcal{T}_{source}$ in Section 4 and the other settings in the Appendix.

For each partition, we train a value classifier, $\mathcal{C}$, in each of the four scenarios shown in Figure 1. These scenarios differ in how the classifier is trained. (1) In the $source$ scenario, $\mathcal{T}_{source}$ is the training set. (2) In the $target$ scenario, $\mathcal{T}_{target}$ is the training set. (3) In the $finetune$ scenario, the classifier is first trained on $\mathcal{T}_{source}$ and then continued to train (i.e., finetuned) on $\mathcal{T}_{target}$. (4) In the all scenario, the training set includes both $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$.

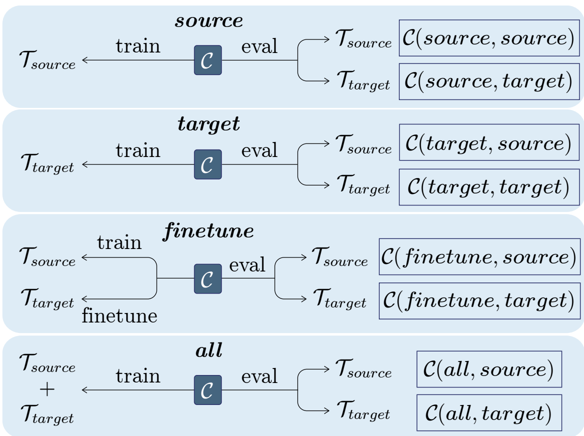

Figure 1: The cross-domain evaluation setting

[1] https://github.com/adondera/transferability-of-values

In each scenario, the classifier is evaluated on both $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$, resulting in eight settings (combinations of training scenario and evaluation set) as shown in Figure 1. For example, $\mathcal{C}(source, target)$ indicates that $\mathcal{C}$ is trained in the $source$ scenario (i.e., on $\mathcal{T}_{source}$) and evaluated on $\mathcal{T}_{target}$.

As we have seven partitions and four scenarios, we train 28 unique models. We evaluate the models on both $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$, covering 56 settings.

3.2 Comparisons

Our experimental setting (partitioning, training scenarios, and evaluation settings) enables a comprehensive cross-domain evaluation of the value classifiers as described below.

Baseline $\mathcal{C}(source, source)$ and $\mathcal{C}(target, target)$ show the performances of a value classifier on the training domain, when no cross-domain training is performed.

Topline $\mathcal{C}(all, source)$ and $\mathcal{C}(all, target)$ represent the ideal scenario, where all data is simultaneously available for training.

Generalizability $\mathcal{C}(source, target)$ and $\mathcal{C}(target, source)$ reflect the ability of a value classifier to generalize to a new domain.

Transferability Comparing $\mathcal{C}(finetune, target)$ and $\mathcal{C}(target, target)$ shows whether the knowledge learned by pretraining on $\mathcal{T}_{source}$ ($finetune$ scenario) has an advantage over the absence of pretraining ($target$ scenario).

Catastrophic Forgetting Comparing $\mathcal{C}(finetune, source)$ and $\mathcal{C}(source, source)$ shows the extent to which the knowledge learned by training on $\mathcal{T}_{source}$ is lost when finetuned on $\mathcal{T}_{target}$.

3.3 Metrics

Since the imbalance in our datasets varies greatly, we report both the micro F_1-score and the macro F_1-score in each setting. The micro F_1-score, m, is the weighted (by class size) mean of the per-label F_1-scores. The macro F_1-score, M, is the unweighted mean of the per-label F_1-scores.

When training and testing on the same set, we use 10-fold cross-validation with fixed splits into training and test data, and report the average F_1-scores over the 10 runs. For consistency, when testing on a set different from the training set, we test on 10 splits of the set (i.e., ultimately on the whole set) and report average F_1-scores.

4 Results and Discussion

We evaluate the performance of the model in four training scenarios ($source$, $target$, $finetune$, all). Table 4 reports the micro and macro F_1-scores of the eight evaluation settings. The columns indicate the dataset used as $\mathcal{T}_{target}$ (e.g., in the BLT column, BLT is $\mathcal{T}_{target}$ and the remaining six datasets compose $\mathcal{T}_{source}$). The final column reports the average F_1-scores over the seven datasets. We also report the results of the majority classifier which labels all tweets as nonmoral (the majority class in all datasets), for both $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$.

We perform Wilcoxon's ranksum test (Hollander and Wolfe, 1999) to evaluate whether two results significantly differ or not. In each column (and in the top-half or the bottom-half), we choose the setting with the highest F_1-score and perform a pair-wise comparison with each of the other settings in that (half) column. We highlight, in bold, the best result and the results that are not significantly different ($p > 0.05$) from the best.

4.1 General Trends

Before cross-domain analysis, we observe some general trends. First, the topline training scenario (all) leads to the best results when evaluating on both $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$ (Table 4). However, all is the ideal scenario. In the top half of the table, $\mathcal{C}(source, source)$ has comparable results to $\mathcal{C}(all, source)$, which is to be expected, since the two models are trained on similar data (six out of seven datasets in the $source$ scenario, all seven in the all scenario). Analogously, in the bottom half of the table, the $\mathcal{C}(finetune, target)$ setting leads to results comparable to $\mathcal{C}(all, target)$. We analyze this result further in Section 4.3.

Second, the results are rather consistent across datasets when evaluating on $\mathcal{T}_{source}$ (top half of Table 4), but have large differences when evaluating on $\mathcal{T}_{target}$ (bottom half of Table 4). These differences can be attributed to BLT and DAV, two highly-imbalanced datasets (Table 3). The class imbalance also justifies the large difference between micro and macro F_1-scores for these two datasets.

4.2 Generalizability

To evaluate generalizability, we analyze the results for the $\mathcal{C}(source, target)$ and $\mathcal{C}(target, source)$ settings. In $\mathcal{C}(source, target)$, $\mathcal{T}_{source}$ includes six datasets and $\mathcal{T}_{target}$ includes one dataset. In contrast, in $\mathcal{C}(target, source)$, $\mathcal{T}_{source}$ includes

Table 4: Results of the four training scenarios evaluated on $\mathcal{T}_{source}$ and $\mathcal{T}_{target}$. The columns indicate the dataset used as $\mathcal{T}_{target}$. We report both micro F_1-score (m, left column) and macro F_1-score (M, right column).

	ALM		**BLT**		**BLM**		**DAV**		**ELE**		**MT**		**SND**		**Average**	
Classifier Setting	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M
$\mathcal{C}(source, source)$	**73.9**	**65.6**	**73.9**	**68.3**	**71.2**	**61.8**	**71.1**	**66.4**	**73.3**	**66.4**	**75.7**	**68.0**	**74.5**	**66.5**	**73.4**	**66.1**
$\mathcal{C}(target, source)$	61.6	37.7	43.8	13.1	62.6	43.0	38.8	5.1	59.3	40.4	52.4	39.1	54.4	36.6	53.3	30.7
$\mathcal{C}(finetune, source)$	70.3	57.2	61.2	47.8	69.2	54.9	56.6	41.9	70.5	61.5	67.7	60.5	68.0	60.8	66.2	54.9
$\mathcal{C}(all, source)$	**73.7**	**65.6**	**73.7**	**68.0**	**71.3**	**62.1**	**71.0**	**66.4**	**73.6**	**66.7**	**75.6**	**67.7**	**74.3**	**66.6**	**73.3**	**66.2**
Majority ($source$)	47.0	6.1	42.3	5.6	49.0	6.2	38.8	5.3	46.1	6.0	49.0	6.2	48.9	6.2	45.9	5.9
$\mathcal{C}(source, target)$	63.7	57.9	63.2	29.2	76.1	75.3	83.9	8.7	63.4	54.8	54.3	51.3	49.2	38.6	64.8	45.1
$\mathcal{C}(target, target)$	**68.0**	56.8	**71.4**	23.5	**84.4**	**84.6**	**92.2**	**9.0**	70.9	52.6	**59.4**	55.9	**65.3**	44.6	**73.1**	46.7
$\mathcal{C}(finetune, target)$	**69.4**	**67.0**	**72.1**	**37.4**	**84.6**	**85.5**	**92.2**	**9.2**	**72.9**	**65.2**	**61.4**	**59.3**	**66.7**	**55.6**	**74.2**	**54.2**
$\mathcal{C}(all, target)$	**69.9**	**67.0**	**71.2**	**34.7**	**83.9**	**85.2**	90.4	**9.3**	71.1	62.3	**61.4**	**59.3**	**66.3**	**55.6**	**73.5**	**53.3**
Majority ($target$)	37.9	5.1	64.8	7.4	28.3	4.2	**92.2**	8.7	44.5	5.7	27.9	4.4	26.4	4.0	46.0	5.6

one dataset and $\mathcal{T}_{target}$ includes six datasets. Thus, $\mathcal{C}(target, source)$ is a more challenging setting for generalization than $\mathcal{C}(source, target)$.

First, we observe that the model achieves better average F_1-scores in the $\mathcal{C}(source, target)$ setting than the majority ($target$) baseline. This indicates that the moral rhetoric learned on a varied array of domains is generalizable to a novel domain to some extent, in spite of the domain-specific nature of moral values. However, the performances in $\mathcal{C}(source, target)$ are not on par with the best results on $\mathcal{T}_{target}$, as we discuss in Section 4.3.

Second, we observe that the model achieves better average F_1-scores in the $\mathcal{C}(target, source)$ setting than the majority ($source$) baseline, despite the more challenging setting. However, the results are just marginally better than the majority ($source$) baseline, showing the difficulty in generalizing from one to multiple domains.

Finally, in both cases, when we look at the results for individual datasets, the generalizability result does not hold for BLT and DAV, which highlights the challenge of generalizing to domains with a skewed distribution of moral values.

4.3 Transferability

Recall that, in the $target$ scenario, a model is only trained on $\mathcal{T}_{target}$, but in the $finetune$ scenario, the model is first trained on $\mathcal{T}_{source}$ and then finetuned on $\mathcal{T}_{target}$. Thus, to evaluate transferability, we compare the $\mathcal{C}(finetune, target)$ and $\mathcal{C}(target, target)$ settings.

From the average F_1-scores in Table 4, we observe that $\mathcal{C}(finetune, target)$ performs better than or on par with $\mathcal{C}(target, target)$—precisely, similar m and 8% increase of M. Thus, the benefits of finetuning are larger for the macro than the micro F_1-scores. This suggests that pretraining on $\mathcal{T}_{source}$, which contains a more varied distribution of labels than $\mathcal{T}_{target}$, improves the prediction of the minority labels in $\mathcal{T}_{target}$.

To transfer knowledge from $\mathcal{T}_{source}$ to $\mathcal{T}_{target}$, typically, we need some labeled data in $\mathcal{T}_{target}$. For the results in Table 4, we used 90% of $\mathcal{T}_{target}$ for training, and the leftover 10% for evaluating at each fold. However, in practice, such a large amount of training data may not be available in the target domain. Thus, we perform an additional experiment to compare $\mathcal{C}(target, target)$ and $\mathcal{C}(finetune, target)$, when trained or finetuned, respectively, on a smaller portion of $\mathcal{T}_{target}$ (10%, 25%, and 50%) and tested on a fixed, randomly selected, 10% of $\mathcal{T}_{target}$. Figure 2 shows this comparison. We report the average results of 10-fold cross-validations performed on each of the seven datasets.

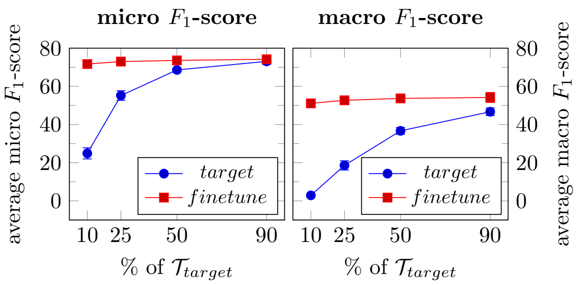

Figure 2: $\mathcal{C}(target, target)$ and $\mathcal{C}(finetune, target)$ results trained with increasing portions of $\mathcal{T}_{target}$

We make an important observation from Figure 2. The finetuning paradigm does not require a large portion of $\mathcal{T}_{target}$ to perform well in the target domain. In contrast, the performance of $\mathcal{C}(target, target)$ increases (but does not surpass

$\mathcal{C}(finetune, target)$) as training data from $\mathcal{T}_{target}$ increases. Indeed, $\mathcal{C}(finetune, target)$ with 10% of $\mathcal{T}_{target}$ performs on par with $\mathcal{C}(target, target)$ trained on 90% of $\mathcal{T}_{target}$. This result shows that transferring the knowledge of values from source domains to a target domain is valuable especially when the target domain has little training data.

4.4 Catastrophic Forgetting

Recall that, in the *source* scenario, a model is only trained on $\mathcal{T}_{source}$, but in the *finetune* scenario, the model is first trained on $\mathcal{T}_{source}$ and then finetuned on $\mathcal{T}_{target}$. Thus, comparing $\mathcal{C}(finetune, source)$ and $\mathcal{C}(source, source)$ provides insight on the extent to which a model forgot about $\mathcal{T}_{source}$ because of finetuning on $\mathcal{T}_{target}$.

We observe that the model suffers from catastrophic forgetting since finetuning on $\mathcal{T}_{target}$ reduces the performance on $\mathcal{T}_{source}$. The forgetting is most evident when finetuning on unbalanced datasets such as DAV than balanced datasets such as BLM. In fact, $\mathcal{C}(finetune, source)$ leads to only slightly worse results than $\mathcal{C}(source, source)$ in BLM (decrease of 2% in m and 7% in M), with the difference being largest in DAV (decrease of 15% in m and 25% in M).

Figure 2 shows that the finetuning paradigm ensures good performances on $\mathcal{T}_{target}$ even when the model is trained on a small portion of $\mathcal{T}_{target}$. Next, we evaluate catastrophic forgetting in the same setting, comparing $\mathcal{C}(source, source)$ and $\mathcal{C}(finetune, source)$ when the model is trained with increasing portions of $\mathcal{T}_{target}$ (10%, 25%, and 50%) as shown in Figure 3.

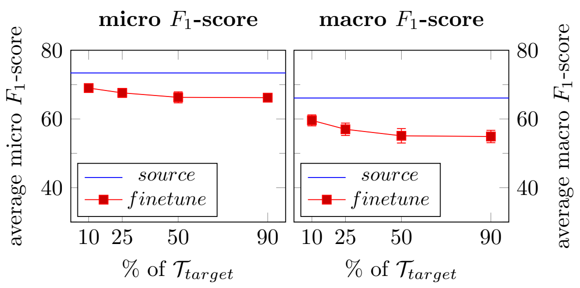

Figure 3: $\mathcal{C}(source, source)$ and $\mathcal{C}(finetune, source)$ results trained with increasing portions of $\mathcal{T}_{target}$

Figure 3 indicates that catastrophic forgetting worsens as the model is trained with a larger portion of $\mathcal{T}_{target}$. $\mathcal{C}(finetune, source)$ trained with 10% of $\mathcal{T}_{target}$ leads to a decrease of 4% in m and 7% in M compared to $\mathcal{C}(source, source)$ (evident by comparing the *source* flat blue line to the first red *finetune* square in Figure 3). Further, $\mathcal{C}(finetune, target)$ trained with 10% of $\mathcal{T}_{target}$ leads to an increase of 7% in m and 6% in M compared to $\mathcal{C}(source, target)$ (evident by comparing the average $\mathcal{C}(source, target)$ in Table 4 to the first red *finetune* square in Figure 2). These results show the tradeoff between the advantage of transfer learning and the impact of forgetting, even when finetuning with a small portion of $\mathcal{T}_{target}$.

4.5 Misclassification Errors

We reported F_1-scores to provide an overview of the model performance in different training settings. Next, we investigate the behavior of the model through the lens of the MFT. We inspect (1) the confusion between morally loaded and nonmoral tweets, and, (2) the mistakes among and within moral foundations since moral foundations are differentially manifested in language (Kennedy et al., 2021). We highlight the following four types of misclassification errors (which add up to 100%):

Error I A tweet labeled with one (or more) values is classified (by the model) as nonmoral.

Error II A tweet labeled as nonmoral is classified with one (or more) values.

Error III A tweet labeled with a value is classified with values from other foundations.

Error IV A tweet labeled as a vice/virtue is classified as the opposite virtue/vice of the foundation.

Table 5 shows the distribution of errors, averaged over the seven datasets.

Table 5: Distribution of errors per setting (in percentage)

Setting	Err. I	Err. II	Err. III	Err. IV
$\mathcal{C}(source, source)$	25.8	34.3	36.3	3.5
$\mathcal{C}(target, source)$	41.8	24.4	32.0	1.8
$\mathcal{C}(finetune, source)$	38.7	27.5	31.3	2.5
$\mathcal{C}(all, source)$	25.9	34.3	36.3	3.4
$\mathcal{C}(source, target)$	34.7	32.3	30.2	2.8
$\mathcal{C}(target, target)$	31.5	27.6	38.5	2.4
$\mathcal{C}(finetune, target)$	36.0	28.6	32.6	2.8
$\mathcal{C}(all, target)$	30.8	33.0	33.1	3.1

Generalizability In $\mathcal{C}(target, source)$, Error I occurs largely more often than the other errors, indicating that, when generalizing from one to several domains, labeling value-laden tweets as nonmoral is the most common mistake. In contrast, in $\mathcal{C}(source, target)$, when generalizing from several to one domain, Error I is less prominent, indicating that the model attempts to classify moral rhetoric in the novel domain.

Transferability Error III is more prevalent in $\mathcal{C}(target, target)$ than $\mathcal{C}(finetune, target)$. Thus, the confusion among moral values reduces when a model is pretrained on the source domain.

Catastrophic Forgetting Error I occurs largely more often in $\mathcal{C}(finetune, source)$ than $\mathcal{C}(source, source)$, indicating that the major type of catastrophic forgetting is missing moral rhetoric in the source dataset.

Finally, Error IV occurs seldom, suggesting that the models generally learn to not confuse between virtues and vices of the same moral foundation.

4.6 Annotators Agreement

We analyze the correspondence between the model predictions and the annotators agreement. Each tweet in the MFTC was annotated by at least three and at most eight different annotators (Hoover et al., 2020, Table 1). More than 99% of the tweets were annotated by three to five annotators and 84% by three or four annotators. As described in Section 2, the majority agreement was selected for training and evaluation—that is, only values annotated by at least 50% of the annotators were retained as correct labels. However, given the subjectivity in value annotation, values labeled by a minority of annotators ought to be considered too.

Tables 6 and 7 show the percentage of annotators that agree with the model predictions considered as errors and accurate, respectively, averaged over the seven datasets. The columns indicate the percentage of annotators agreeing with the model prediction. For instance, if one out of the four workers who annotated a tweet agrees with the model prediction, we record a 25% agreement.

Table 6: Distribution (in percentage) of classification errors and annotators agreement percentage

Setting	$\mathbf{0}$	$(\mathbf{0}, \mathbf{25}]$	$(\mathbf{25}, \mathbf{34}]$	$(\mathbf{34}, \mathbf{50})$
$\mathcal{C}(source, source)$	26.1	22.3	45.0	6.6
$\mathcal{C}(target, source)$	49.5	18.0	28.5	3.9
$\mathcal{C}(finetune, source)$	38.5	20.2	36.1	5.2
$\mathcal{C}(all, source)$	26.3	22.2	45.0	6.5
$\mathcal{C}(source, target)$	40.2	23.2	30.4	6.2
$\mathcal{C}(target, target)$	19.7	30.7	40.6	8.9
$\mathcal{C}(finetune, target)$	21.2	30.5	39.9	8.4
$\mathcal{C}(all, target)$	25.6	27.5	39.0	7.9

First, we analyze the classification errors in Table 6. We observe that the sum of the last three columns is always larger than 50%. This indicates that, in all settings, more than half of the model classification errors are not severe in that at least one human annotator agrees with the model prediction. Then, we notice that the settings with the highest incidence of 'bad' classification errors (i.e., where no annotators agree with the model prediction) are those employed to evaluate generalizability ($\mathcal{C}(target, source)$ and $\mathcal{C}(source, target)$) and catastrophic forgetting ($\mathcal{C}(finetune, source)$). These results are explained by the harder challenge represented in these settings (refer to Sections 4.2 and 4.4 for a more in-depth discussion). Finally, we observe that there is a small percentage of errors with agreement between 34% and 50%. For the agreement to be in this range, a tweet must have been annotated by at least 5 annotators. However, 84% of the tweets in the MFTC have been annotated by four annotators or less, thus resulting in a smaller agreement in the last column.

Table 7: Distribution (in percentage) of correct predictions and annotators agreement percentage

Setting	$[\mathbf{50}, \mathbf{66})$	$[\mathbf{66}, \mathbf{75})$	$[\mathbf{75}, \mathbf{100})$	$\mathbf{100}$
$\mathcal{C}(source, source)$	16.9	24.4	20.9	37.7
$\mathcal{C}(target, source)$	16.8	20.0	20.2	43.1
$\mathcal{C}(finetune, source)$	17.0	22.7	20.9	39.4
$\mathcal{C}(all, source)$	17.0	24.5	20.9	37.7
$\mathcal{C}(source, target)$	15.0	27.5	18.5	39.0
$\mathcal{C}(target, target)$	15.0	27.7	18.8	38.5
$\mathcal{C}(finetune, target)$	15.8	28.5	18.7	37.0
$\mathcal{C}(all, target)$	15.7	28.4	18.8	37.2

Second, we analyze the correct predictions in Table 7. We notice, in all settings, a high correspondence between 100% agreement among annotators and correct model predictions—that is, tweets annotated with consistent agreement reliably lead to correct predictions. Further, we observe that the distributions of agreement and correct predictions are consistent across different settings.

5 Related Work

We review closely related works on value estimation from text, and on cross-domain classification in NLP subfields relevant to value classification.

5.1 Value Estimation from Text

Value estimation has been addressed from both unsupervised and supervised approaches. Unsupervised methods exploit value lexicons to identify values in text. Value lexicons are generated manually (Graham et al., 2009), via semi-automated methods (Wilson et al., 2018; Rezapour et al., 2019; Araque et al., 2020; Hopp et al., 2021), or expanded from an initial seed via NLP techniques (Ponizovskiy et al., 2020; Araque et al., 2021). Value lexicons are used to identify values in text through word count software (Pennebaker et al., 2001) or similarity in embedding space (Garten et al., 2018; Shen et al., 2019; Bahgat et al., 2020). However, adapting a lexicon to a novel domain is a significant additional effort as it requires identifying words that are relevant and removing words that are not relevant in the novel domain.

Supervised methods employ the classification paradigm (Lin et al., 2018; Mooijman et al., 2018; Hoover et al., 2020; Alshomary et al., 2022; Kiesel et al., 2022). A textual dataset is annotated with values belonging to a value taxonomy, and the labels are used to train a supervised model. This approach is akin to the one we use in this paper. However, in the reviewed literature, no emphasis is put on the effect of cross-domain training. Further, several of the works mentioned above (Lin et al., 2018; Mooijman et al., 2018; Hoover et al., 2020) use binary classification to independently predict the presence of a value in text. That is, given N values, N classifiers are employed (one per value). However, it has been shown that modeling relationships among values (and additional contextualizing information such as actors) helps improve downstream performances (Johnson and Goldwasser, 2018; Roy et al., 2021). Thus, we train a multi-label value classifier, similarly to Alshomary et al. (2022) and Kiesel et al. (2022). Furthermore, our objective is not to compare binary and multi-label value classification but to evaluate the cross-domain capabilities (generalizability, transferability, and catastrophic forgetting) of a multi-label value classifier.

5.2 Datasets with Moral Content

The recent success of NLP models has sparked a surge of research in constructs akin to moral values, e.g., moral norms, ethical judgments, and social biases. Researchers have collected large datasets annotated with the related implicit components of human language similar to the MFTC (Section 2). Forbes et al. (2020) introduced SOCIAL-CHEM-101, a corpus of almost 300,000 rules-of-thumb aimed at learning social and moral norms. Sap et al. (2020) collected the Social Bias Inference Corpus with the intent of modeling the way in which people project social biases onto each others. Hendrycks et al. (2021) proposed the ETHICS dataset to assess basic knowledge of ethics through well-studied theories of normative ethics (such as deontology and utilitarianism). Lourie et al. (2021) introduced SCRUPLES, a dataset composed of 625,000 ethical judgments over 32,000 real-life anecdotes. Finally, Emelin et al. (2021) presented *Moral Stories*, a crowd-sourced collection of contextualized narratives with the intent of investigating grounded, goal-oriented social reasoning.

These datasets offer an unprecedented opportunity for studying the social and moral aspects of language. In our research we employ the MFTC as the same moral value theory is used to annotate data in seven different domains, allowing for a direct cross-domain comparison.

5.3 Cross-Domain NLP Classification

Cross-domain classification is gaining attention (Aji et al., 2020; Nguyen et al., 2021; Rongali et al., 2021; Bornea et al., 2021; Markov and Daelemans, 2021). Ruder (2019) provides an overview of the basic terminology, including generalizability, transferability, and catastrophic forgetting.

Cross-domain classification has been investigated in NLP tasks such as sentiment analysis (Al-Moslmi et al., 2017; Qu et al., 2019; Du et al., 2020), fake news detection (Fung et al., 2021; Silva et al., 2021; Yuan et al., 2021), and argument mining (Al-Khatib et al., 2016; Daxenberger et al., 2017; Thorn Jakobsen et al., 2021). These tasks are similar to value classification in that they aim to classify high-level constructs (such as sentiments and arguments). However, value classification stands out for its multi-label and domain-specific nature. Also, cross-domain classification is particularly important for values because reasoning about values (Pommeranz et al., 2012) and generating value-annotated datasets is very difficult.

6 Conclusions and Directions

We perform a comprehensive cross-domain evaluation of a multi-label value classifier, by comparing a deep learning model (BERT) in seven domains with four cross-domain training scenarios. Our aim is to support practical applications of moral rhetoric classification, e.g., the detection of radicalism through the study of moral homogeneity (Atari et al., 2021), the prediction of violent protests (Mooijman et al., 2018), the identification of moral concerns of citizens (Mouter et al., 2021; Siebert et al., 2022), and the extraction of moral rhetoric supporting both stances and arguments (Draws et al., 2022; van der Meer et al., 2022). Our findings inform both computer scientists and social scientists on training value classifiers. However, we do not provide a fixed recipe since the right model and approach depend on the time, resources, and data available.

We show that a value classifier generally exhibits the ability to classify moral values across domains. However, the results are highly dependent on the distribution of moral rhetoric in a domain.

Our experiments support the following key findings. First, a value classifier can generalize to novel domains, especially when trained on multiple domains. However, its performance on the novel domain improves even when trained with a small portion of data from the novel domain. Second, pretraining a value classifier with data from different domains has three benefits when finetuning the classifier. It yields (1) better performances on the novel domain than other settings, (2) good performances even when little training data is available in the novel domain, and (3) smaller confusion among moral values, especially among those less frequent in the novel domain. Third, finetuning on a novel domain causes catastrophic forgetting of the domain it was pretrained with, even when finetuning on a small portion of data from the novel domain. Thus, the tradeoff between benefits of transferability and adverse effects of forgetting must be considered in choosing the extent of finetuning. Finally, despite the challenging nature of cross-domain value classification, the majority of classification errors are not severe in that, in all evaluation settings, at least one annotator agrees with the model prediction.

Our investigation opens avenues for additional experiments with advanced methods to improve transfer learning (Howard and Ruder, 2018; Jiang et al., 2020; Nguyen et al., 2021) and mitigate catastrophic forgetting (Kirkpatrick et al., 2017; Li and Hoiem, 2018; Thompson et al., 2019). Further, based on the analysis of classification errors, we suggest incorporating the annotators (dis-) agreement into the training of the model, e.g., by employing the full distributions of annotations, as opposed to the current majority approach (Uma et al., 2021).

7 Ethical Considerations

We discuss three ethical considerations relevant to our work. First, the MFTC is composed of monolingual tweets about US-centric topics. Whether or not our conclusions hold for results across different languages and cultures is yet to be evaluated. This limitation may cause the perpetuation of Western biases and values (Mehrabi et al., 2021). However, we believe that our experimental setup offers a systematic approach to studying such cultural influences when pertinent data is available.

Second, the MFTC has low annotator agreement (Hoover et al., 2020, Table 6), potentially caused by the subjectivity and complexity of annotating values. Selecting the majority label as golden label may perpetuate the 'tyranny' of the majority, which is especially dangerous when dealing with values. We expose the impact of the annotator agreement in Section 4.6 and identify an avenue for addressing it as a future direction in Section 6.

Finally, the importance of understanding moral values has been recognized by computer scientists (Russell et al., 2015) and designers (Friedman et al., 2008). However, we recognize that value classification can be misused, especially, when sensitive attributes such as gender and race are attached to the data. For instance, authorities could use it to automatically identify and suppress liberal minorities in non-liberal countries. Additional research is necessary for addressing such problems, e.g., by devising techniques that mitigate bias and unfairness by design (Kleinberg et al., 2018; Dinan et al., 2020; Vargas and Cotterell, 2020).

Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research. Furthermore, we thank Florentin Arsene for his contribution in previous iterations of the project.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In Neural Machine Translation, What Does Transfer Learning Transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7701–7710, Online. Association for Computational Linguistics.

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th Conference* *on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 16–24, Auckland. IFAAMAS.

Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn J. M. Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8):18–28.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '16, pages 1395–1404.

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 1–16, Dublin, Ireland. Association for Computational Linguistics.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction. *Knowledge-Based Systems*, 191:1–29.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2021. The Language of Liberty: A preliminary study. In *Companion Proceedings of the Web Conference 2021*, WWW '21 Companion, pages 1–4, Ljubljana, Slovenia. Association for Computing Machinery.

Mohammad Atari, Aida Mostafazadeh Davani, Drew Kogon, Brendan Kennedy, Nripsuta Ani Saxena, Ian Anderson, and Morteza Dehghani. 2021. Morally Homogeneous Networks and Radicalism. *Social Psychological and Personality Science*, 12:1–11.

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.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual Transfer Learning for QA Using Translation as Data Augmentation. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 12583–12591, Online.

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.

Kinzang Chhogyal, Abhaya Nayak, Aditya Ghose, and Hoa K. Dam. 2019. A Value-Based Trust Assessment Model for Multi-Agent Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI '19, pages 194–200.

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.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2055–2066.

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 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '19, page 4171–4186.

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.

Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 135–145, Regensburg, Germany. Association for Computing Machinery.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 4019–4028.

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.

Batya Friedman, Peter H. Kahn, and Alan Borning. 2008. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*, pages 69–101. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, Online. Association for Computational Linguistics.

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

Justin Garten, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis: Distributed Dictionary Representation. *Behavior Research Methods*, 50(1):344–361.

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.

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.

Myles Hollander and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. Wiley, New York, USA.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text. *Behavior Research Methods*, 53:232–246.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-Tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 328–339.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and Efficient Fine-Tuning for Pretrained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 2177–2190. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2018. Classification of Moral Foundations in Microblog Political Discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

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 1–13, Dublin, Ireland. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27.

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

Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. 2009. Values as Lived Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1141–1150. ACM Press.

Zhizhong Li and Derek Hoiem. 2018. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring Background Knowledge to Improve Moral Value Prediction. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, pages 552–559, Barcelona, Spain. IEEE.

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

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

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13470–13479.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

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).

Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? The Effect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 777—-783. Association for Computing Machinery.

Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, pages 907–915, Online. IFAAMAS.

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.

Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn J. M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1706–1710, Auckland. IFAAMAS.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 5414–5426, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic Inquiry and Word Count (LIWC). *Mahway: Lawrence Erlbaum Associates*, 71.

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.

Xiaoye Qu, Zhikang Zou, Yu Cheng, Yang Yang, and Pan Zhou. 2019. Adversarial category alignment network for cross-domain sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '19, pages 2496–2508, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the Measurement of Social Effects by Capturing Morality. In *Proceedings of the 10th*

Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 35–45, Minneapolis, Minnesota, USA.

Subendhu Rongali, Beiye Liu, Liwei Cai, Konstantine Arkoudas, Chengwei Su, and Wael Hamza. 2021. Exploring Transfer Learning For End-to-End Spoken Language Understanding. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 13754–13761, Online.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying Morality Frames in Political Tweets using Relational Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 9939–9958. Association for Computational Linguistics.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, NUI Galway.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 5477–5490, Online. Association for Computational Linguistics.

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

Marc Serramia, Maite López-Sánchez, Stefano Moretti, and Juan Antonio Rodríguez-Aguilar. 2021. On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems*, 35(2):1–38.

Yiting Shen, Steven R. Wilson, and Rada Mihalcea. 2019. Measuring Personal Values in Cross-Cultural User-Generated Content. In *Proceedings of the 10th International Conference on Social Informatics*, SocInfo '19, pages 143–156. Springer.

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 *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence*, HHAI '22, pages 1–14, Amsterdam, the Netherlands. IOS Press.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21, pages 557–565.

Nate Soares and Benya Fallenstein. 2017. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity: Managing the Journey*, pages 103–125. Springer, Berlin.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '19, pages 2062–2068, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of the Tenth Joint Conference on Lexical and Computational Semantics*, *SEM 2021, pages 263–277, Online. Association for Computational Linguistics.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening Between the Lines: Learning Personal Attributes from Conversations. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 1818–1828.

Alexandra N Uma, Dirk Hovy, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence*, HHAI '22, pages 1–15, Amsterdam, the Netherlands. IOS Press.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, NeurIPS '17, pages 5998–6008, Long Beach, CA, USA.

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.

Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. 2021. Improving Fake News Detection with Domain-Adversarial and Graph-Attention Neural Network. *Decision Support Systems*, 53:113633.