Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

A Data Fusion Framework for Multi-Domain Morality Learning

Siyi Guo, Negar Mokhberian, Kristina Lerman

Information Sciences Institute, University of Southern California
Marina del Rey, CA, USA
{siyiguo,nmokhber,lerman}@isi.edu

Abstract

Language models can be trained to recognize the moral sentiment of text, creating new opportunities to study the role of morality in human life. As interest in language and morality has grown, several ground truth datasets with moral annotations have been released. However, these datasets vary in the method of data collection, domain, topics, instructions for annotators, etc. Simply aggregating such heterogeneous datasets during training can yield models that fail to generalize well. We describe a data fusion framework for training on multiple heterogeneous datasets that improve performance and generalizability. The model uses domain adversarial training to align the datasets in feature space and a weighted loss function to deal with label shift. We show that the proposed framework achieves state-of-the-art performance in different datasets compared to prior works in morality inference.

Introduction

Morality helps people distinguish between "right" and "wrong" and governs their everyday behaviors and interactions with others (Hofmann et al. 2014). Morality also shapes judgments, attitudes and beliefs, creating differences in the moral experience of individuals across cultural groups (Haidt, Joseph et al. 2007). Studies have linked moral sentiment to partisan ideologies (Graham, Haidt, and Nosek 2009), messaging strategies in politics (Wang and Inbar 2021) and news (Mokhberian et al. 2020), and even real-world violence (Mooijman et al. 2018).

Researchers have developed a scale to quantify morality, which represents people's intuitive ethical reactions to social dilemmas. The so-called Moral Foundations Theory (MFT) (Haidt and Joseph 2004; Graham et al. 2013) characterizes morality along five dimensions:

- *Care/Harm*: dislike of suffering.
- *Fairness/Cheating*: proportionality, justice and rights.
- *Loyalty/Betrayal*: attachment to one's identified group.
- *Authority/Subversion*: respect for authority and tradition.
- *Sanctity/Degradation*: endorsement of purity and cleanliness, avoidance of pollution and decay.

While early research in morality science relied on moral foundation questionnaires and vignettes to characterize morality along these dimensions (Graham et al. 2011), recent works have begun to infer morality from text. Automated natural language processing (NLP) methods (Fulgoni et al. 2016; Garten et al. 2018; Lin et al. 2018; Xie, Hirst, and Xu 2020) have enabled researchers to scale up moral foundation inference to large text corpora of news articles and messages posted on social media, opening new avenues for studying morality.

The more sophisticated approaches train language models on ground truth datasets of text labeled by human annotators according to its moral expression. The trained models are then used to recognize morality of a new text. The rapid growth of interest in language and morality has produced several ground truth datasets for moral foundation inference (Hoover et al. 2020; Johnson and Goldwasser 2018; Forbes et al. 2020; Hopp et al. 2021; Trager et al. 2022). Researchers hope that training models on multiple datasets will yield better performance and generalizability. However, the labeled datasets are *heterogeneous*: they vary by domain (news vs social media), topics covered (politics vs health), task granularity (label the moral foundations vs its vices and virtues), annotator population (few experts vs a crowd), annotation instructions (select text expressing a given moral foundation vs select relevant moral foundations for a given text), etc. Blindly combining heterogeneous data during training may bias predictions (Bareinboim and Pearl 2016). For example, models trained on aggregated data may give predictions that are inconsistent with predictions of models trained on disaggregated data, a phenomenon called Simpson's paradox (Alipourfard, Burghardt, and Lerman 2021).

We address the challenge of training moral foundation classifiers on heterogeneous datasets via a data fusion-inspired approach. Unlike training a classifier on aggregated data, our proposed approach uses domain adversarial training (Ganin and Lempitsky 2015) to map the features (text embeddings) from different datasets onto a common embedding space, thus mitigating the problem brought by the heterogeneity of the texts and topics in the datasets and improves the generalizability across different data sources.

In addition to differences in the feature space, datasets can also vary in the distribution of labels. For example, tweets related to the pandemic have many more messages express-

ing the *care/harm* foundation than tweets related to civil unrest. Unless accounted for, differences in label distribution will hurt classifier performance. To mitigate this problem, we propose to use a weighted loss function that balances among different label classes and between positive and negative data examples.

Compared to previously used morality detection methods, our proposed framework achieves state-of-the-art performance on many datasets in out-of-domain testing. We believe that our work is the first one improving the generalizability of models for moral foundation inference with multi-dataset learning and a domain adaptation approach.

Related Works

Morality Inference Many prior works have contributed to the development of methods to classify moral foundations from text. One type of methods is dictionary-based, which relies on the use of lexical resources, such as the Moral Foundations Dictionary (MFD) (Haidt and Joseph 2004; Haidt and Graham 2007). Some researchers proposed to use distributed dictionary representations (DDR) to capture the semantic similarities between words (Garten et al. 2018). Another method measured the distance of a text from the axes defined by the words representing the virtues and vices of the moral foundations in the embedding space (Mokhberian et al. 2020).

With recent advances in transformer-based language models such as BERT (Devlin et al. 2019), researchers have shown that these large pre-trained language models have a sense for social norms. Schramowski et al. (2022) demonstrated that BERT captures general morality and identifies right or wrong actions. Some studies have leveraged these language models to calculate contextualized embeddings for moral foundations inference (Xie, Hirst, and Xu 2020; Kennedy et al. 2021; Hofmann et al. 2022). Other researchers have made effort in gathering and annotating datasets for moral foundations analysis (Hoover et al. 2020; Trager et al. 2022).

As the interest in studying morality grows, especially on large quantities of social media data, inferring morality on large unlabeled data becomes a common challenge. Thus, how to utilize limited resources and perform out-of-domain inference becomes an important research question. Islam and Goldwasser (2022) proposed a minimally supervised framework by learning on a combination of many weak labels and a smaller amount of gold labels to analyze morality related to the COVID vaccine topic. Pacheco et al. (2022) also studied morality detection in a low-resource setting to analyze COVID-19 vaccine debates. Their model is trained on a larger quantity of available out-of-domain labeled data and finetuned on a small amount of in-domain labels.

However, in-domain gold labels for targeted data are not always available. Annotating morality is an especially challenging task, suffering from great time complexity and low annotator agreement. In this work, we explore a more challenging setting, performing out-of-domain morality inference without any available in-domain labels. We propose to utilize already available heterogeneous datasets with morality labels, and adopt domain adaptation ideas to improve model's out-of-domain performance.

Domain Adaptation The unsupervised domain adaptation problem is an active research area in machine learning. Researchers have developed various approaches, including structural correspondence learning (Blitzer, McDonald, and Pereira 2006), joint distribution matching using max mean discrepancy (Long et al. 2013) and mixture of experts (Guo, Shah, and Barzilay 2018). Other recent works have tried to improve out-of-domain prediction by improving the training data quality. Le Bras et al. (2020) studied AFLite, a method that cleans the redundant information in training data thus mitigates model's dependence on spurious correlations.

Another popular method is the domain adversarial neural network (DANN) (Ganin and Lempitsky 2015). On top of a regular feature extractor and a label classifier, this model structure includes a domain classifier, an adversary that pushes the feature extractor to generate domain-invariant features, and thus facilitates domain adaptation. This architecture is shown to be successful in other NLP tasks such as aspect-dependent text classification (Zhang, Barzilay, and Jaakkola 2017) and stance detection (Allaway, Srikanth, and McKeown 2021; Hardalov et al. 2021). DANN is a great method to deal with feature distribution shift, which is one of the major problems in merging heterogeneous datasets with different topics and annotation processes. Thus, we incorporate the DANN structure as a part of our model.

The similar idea of merging multiple datasets and incorporating domain adaptation techniques has been tested out in prior works for stance detection (Hardalov et al. 2021; Li, Zhao, and Caragea 2021). However, stance detection is very different from morality detection, as their main challenges are the varying label sets (e.g. for/against in one dataset, comment/support/query/deny in another dataset) and varying target sets (e.g. Trump, climate change, legalization of abortion, etc). These prior works have thus put more focus on adapting among different label and target sets. Morality prediction, on the other hand, is backed with the well-defined Moral Foundations Theory and has the same set of labels for most of our datasets. Instead, our major challenge is the feature and label distribution shift among datasets. On top of merging datasets and using domain adversarial training, we propose to use a weighted loss function to balance between different label classes.

Methods

We propose a Domain Adapting Moral Foundation inference model (**DAMF**) for fusing annotated data from multiple domains. The model (Figure 1) has five parts. BERT encodes texts into contextualized embeddings. The transformation module facilitates the text embeddings to be more domain-invariant. The moral foundation classifier with a weighted loss function performs the main morality detection task. The domain classifier works as the adversary, pushing the BERT encoder to learn domain-invariant embeddings to fool the domain classifier. Additionally, the reconstruction module attached to the BERT encoder avoids the model from being corrupted by adversarial training. The objective of the

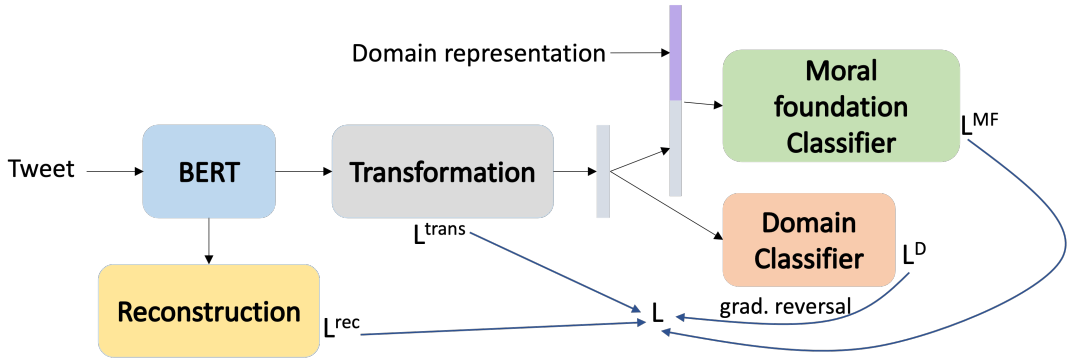

Figure 1: Model structure of **DAMF**: BERT working as text encoder, a transformation module facilitating the text embedding to be more domain-invariant, a moral foundation classifier with a weighted loss function, an adversarial domain classifier, and a reconstruction module to avoid the model being corrupted by adversarial training. The objective is to learn domain-invariant embeddings that can fool the domain classifier and also performs well in morality detection.

model is to learn domain-invariant embeddings to allow heterogeneous domains to align in the embedding space, and also to succeed in morality detection.

Text Encoder We use the pre-trained, transformer-based language model BERT to transform input messages into embedding vectors.

Domain-invariant Transformation The output embeddings from the text encoder contains domain-specific information. To facilitate the transfer across different domains, we follow the idea from Zhang et al. (2017) and add a linear domain-invariant transformation layer:

$$x^{trans} = W^{trans} x^{BERT}$$

where W^{trans} is the weight for the linear layer, x^{BERT} is the embedding generated from the BERT text encoder, and x^{trans} is the embedding after this transformation layer. This transformation is regularized to the identity I to ensure text information is still retained and won't be wiped out by the adversary. Therefore, from this step we obtain a loss term L^{trans}:

$$L^{trans} = ||W^{trans} - I||^2$$

Moral Foundation Classifier This module is in charge of learning the primary task: classifying moral foundations. It has a linear layer, a ReLu activation, a dropout layer, and a second linear layer. The input to this classifier is the embeddings generated by the previous steps (x^{trans}) concatenated to a one-hot domain embedding. This is to facilitate the model learning the relationship between the input tweet and its domain.

One tweet can be related to multiple moral foundations. For example, this tweet during the COVID-19 outbreak "I'm tired of people and media talking about all the merchandise that been looted, what about the minority lives that been lost" expresses both harm and cheating moral foundations. Therefore, we formulate moral foundation inference as a multi-label classification task. There are 10 classes in total (e.g. care, harm, etc) and each class itself is a binary prediction.

Weighted Loss Function One problem in moral foundation inference is that the number of positive and negative examples is severely imbalanced. That is, we have many more non-moral data than those with morality labels. Moreover, the number of examples associated with different moral foundation classes can also vary greatly (Figure 4). To balance between positive and negative examples, as well as among different classes, we use a weighted binary cross-entropy loss L^{MF} combined with a sigmoid layer in the moral foundation classifier:

$$w_c = \frac{\#\ negative\ examples\ in\ c}{\#\ positive\ examples\ in\ c}$$

$$L^{MF} = \frac{1}{N}\sum_{n=1}^{N} -w_c \cdot y^c \cdot log\sigma(x_n^c) + (1-y^c) \cdot log(1-\sigma(x_n^c))$$

where c is the class a data example belongs to (eg. care), w_c is the weight for this class, n is the data example index, x_n^c is the final-layer embedding of the n^{th} example which belongs to the class c, $\sigma()$ represents the sigmoid function, and y^c is the label.

Domain Classifier This is the adversary in the model. It distinguishes which domain an input example comes from and pushes the BERT encoder to produce domain-invariant embeddings that can fool the domain classifier itself. The domain classifier has the same feed-forward network structure as the moral foundation classifier, but a different loss L^D - a cross-entropy loss with a softmax layer. This loss is to be *maximized* during the adversarial training (our goal is to fool this domain classifier), which is achieved by connecting the domain classifier to the other parts of the model with a gradient reversal layer (Ganin et al. 2016).

Reconstruction Module To compete with the adversarial domain classifier, the BERT encoder wants to wipe out as much domain-specific information as possible. However, if not controlled well, the BERT encoder could be overly corrupted, and too much information would be lost in the representations learned. Therefore, we add a reconstruction module to ensure the BERT encoder can still generate representations close to the original contextual BERT embeddings. The

reconstruction module has a linear layer followed by a tanh activation. Its loss L^{rec} is calculated as the mean squared error between the reconstructed embeddings and the original embeddings generated by BERT with no domain adversarial training:

$$L^{rec} = ||tanh(Wx + b) - tanh(x_{orig})||^2$$

where x is the embeddings generated by the current model, W and b are the weight and bias for a linear transformation on x, and x_{orig} is the embeddings generated by BERT model with no domain adversarial training.

The overall loss function of the whole model is:

$$L = \lambda^{rec} \cdot L^{rec} + \lambda^{trans} \cdot L^{trans} + L^{MF} - \lambda^{D} \cdot L^{D}$$

where λ^{rec}, λ^{trans} and λ^{D} are hyper-parameters to be tuned while training. The first three terms in this loss are *minimized* with respect to the model parameters, whereas the last loss term for the domain classifier adversary is *maximized* with respect to its parameters. This is to achieve the objective of fooling the domain classifier and generating domain-invariant embeddings. During training, the maximization of the domain classifier loss term is achieved by using a gradient reversal layer (Ganin et al. 2016).

Experiments

Data

We evaluated the performance of the proposed model on labeled moral foundations data in the out-of-domain setting.

Moral Foundations Twitter Corpus (MFTC) This dataset (Hoover et al. 2020) includes 35K English tweets, and is labeled by trained human annotators with moral foundation labels in 11 classes (eg "care," "harm," "fairness," etc), including a "non-moral" class. The tweets cover six different topics: (1) Black lives matter (BLM) civil protests, (2) All Lives Matter (ALM), (3) the Baltimore protest against the death of Freddie Grey, (4) the 2016 USA Presidential election, (5) Hurricane Sandy and (6) hate speech. We follow the standard practice and aggregate the annotations from multiple annotators using majority vote as the true label. We discard tweets that don't reach a majority vote.

Covid This dataset (Rojecki, Zheleva, and Levine 2021) contains 2,648 tweets related to the COVID19 pandemic. These tweets were manually annotated with 10 moral foundation classes.

Congress This dataset is a collection of 2,050 English-language tweets by members of US Congress (Johnson and Goldwasser 2018). The dataset covers six different political topics: (1) abortion, (2) the Affordable Care Act, (3) guns, (4) immigration, (5) LGBTQ rights, and (6) terrorism.

The extended Moral Foundations Dictionary (eMFD) This dataset differs in terms of source of text and the annotation process. It contains 2995 English news articles on a variety of topics. The labeling task was crowd-sourced to 557 annotators, who were asked to mark all text segments expressing a given moral foundation in 15 randomly selected news articles (Hopp et al. 2021). After processing, this dataset contains 36K labeled examples.

Data	Type	Size
MFTC	tweets	18991
Covid	tweets	2430
Congress	tweets	1849
eMFD	news articles	35985

Table 1: An overview of the datasets

Table 1 shows the sizes of the four datasets after preprocessing. Figure 2 demonstrates the distinct topics discussed in these data by word clouds. We process the texts in all datasets by removing URLs, replacing all mentions with "@user", removing hashtags, replacing emojis with their text descriptions[1], and removing all non-ASCII characters. We took the maximum sequence length to be 50 when training with BERT and DAMF, because the majority of the datasets are tweets or segments in news articles with short lengths. Each data example has binary labels in 10 moral classes, and a non-moral example would have "False" for all 10 classes.

We test different models on four datasets separately. For each test dataset, we experiment with training on a single dataset or combinations of multiple labeled datasets.

Model Training

We compare our proposed model **DAMF** with three baseline models, (1) Distributed Dictionary Representations (DDR) (Garten et al. 2018), (2) BERT (Devlin et al. 2019) and (3) BERT improved with the Lightweight Adversarial Filtering (AFLite) method (Le Bras et al. 2020).

To evaluate the models, we use F1 score weighted by the number of true examples for each class. Each experiment is repeated with 5 random seeds to calculate the mean and the standard deviation.

Distributed Dictionary Representations (DDR) DDR (Garten et al. 2018) is an unsupervised lexicon-based method built on the Moral Foundations Dictionary (MFD), and it is a classic method widely used in prior works (Garten et al. 2016; Hoover et al. 2020; Trager et al. 2022). It represents each moral foundation class by the centroid of the embedding vectors of all words in this category of MFD, and then calculates the cosine similarity between an input text (as the average of its word embeddings) with each of the moral foundation class embedding. Here we use word2vec word embeddings (Mikolov et al. 2013). The moral foundation class label is predicted as the one with the largest similarity score to the input text.

BERT BERT is used in prior works to detect morality (Roy and Goldwasser 2021) and is *the current state-of-the-art method for morality prediction* (Trager et al. 2022). As a

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

Figure 2: Word clouds for each dataset.

supervised model, BERT can perform representation learning with respect to specific tasks or datasets and thus improves the performance. Therefore, finetuning BERT offers us a strong baseline.

We finetune BERT with a linear prediction layer attached. To mitigate over-fitting, we implement a dropout layer with a rate of 0.3. For multi-label prediction, we use the binary cross-entropy loss with sigmoid function. After the naive dataset merge is done, we split the training dataset into 80-20 train-validation sets, and we predict on separate test data. We train different datasets with 20 epochs and stop early when the model achieves the best validation F1 score. The batch size is 64. We use the Adam optimizer with an initial learning rate of 5e-5, and we use a scheduler $lr = lr_{init}/((1+\alpha \cdot p)^{\beta})$, where $p = \frac{current\ epoch}{total\ epoch}$ and we set $\alpha = 10$ and $\beta = 0.25$.

BERT+AFLite To have an even stronger baseline incorporating domain adaptation techniques, we select BERT combined with the Lightweight Adversarial Filtering (BERT+AFLite). AFLite filters out the data points that are too easy to be predicted by weak classifiers and minimizes the ability of a model to exploit spurious correlations, thus improves model's out-of-domain performance (Sakaguchi et al. 2021; Le Bras et al. 2020). We use a single linear layer of neural network as the weak classifier, iteratively train and test on different random partitions of the data, and discard texts if the linear classifier can predict their moral foundation using vanilla BERT embedding. We perform iterative filtering until the data size is shrunk to 50% or until the data size changes from the last round by less than 2%, whichever is achieved first. AFLite filtered out *MFTC* by 50%, the *covid* data by 10%, the *congress* data by 15%, and *eMFD* by 28%. Then, we train a BERT model with the filtered data using the same procedure described in the last section.

DAMF We train the DAMF model in a semi-supervised fashion. That is, in addition to the usual labeled training data, we also use unlabeled data from the target domain in each training data batch. This way, the domain adapting module in our model can align the feature distributions of different datasets including the target data. The data from different domains in each batch are balanced in size. The morality labeled training data passes through both the moral foundation classifier and the domain classifier, and the unlabeled data in the target domain only passes through the domain classifier. We then evaluate on a separate hold-out test dataset.

We train different datasets with 60 epochs in total. During the first 15 epochs, the model is trained without the domain adversarial module activated. This is for the model to better learn the morality labels first. The batch size is 64. We use the Adam optimizer. The learning rate and the scheduler are the same as for BERT model.

When training with a domain adversary, it is challenging to control its power. If too powerful, it easily wipes out useful information in the feature embeddings and leads to bad moral foundation predictions. We control the hyperparamters λ^{rec}, λ^{trans} and λ^{D}. Following Ganin et al. (2015) and Allaway et al. (2021), we control λ^{D} indirectly by a parameter γ:

$$\lambda^{D} = 2/(1+e^{-\gamma \cdot p}) - 1$$

where $p = \frac{current\ epoch - epochs\ trained\ w/o\ adversary}{total\ epochs}$. Thus, we perform hyperparameter search on λ^{rec} in $[0, 0.1, 0.5, 1]$, λ^{trans} in $[no\ trans, 0.01, 0.1, 1, 10]$ and γ in $[0.1, 1, 10]$ (see Table 5 in Appendix).

Results

Heterogeneity in Feature and Label Spaces

We first examine the heterogeneity of different datasets and show the two critical aspects that need to be considered during multi-domain learning: how datasets differ in the feature space and how they differ in label space. In the feature space (Fig. 3), different datasets have different distributions of embeddings. For example, the *congress* dataset, which focuses on messages on US political issues, only takes a subset in the TSNE space compared to the *MFTC* dataset, which includes messages posted on social media on six various topics. Another example *eMFD* contains text of news articles and is also annotated by a crowd. It is very different from other datasets, which have been annotated by a few experts. In Fig. 3, *eMFD* shows a more distinct distribution compared to all other Twitter-based datasets. We therefore expect that a model trained on *eMFD* would not generalize well to the other datasets (Table 2 row 3, row 9, row 16, BERT model performance).

In addition to features, the label distributions among heterogeneous datasets can also differ dramatically (Fig. 4). This is because people express very different moral concerns on different topics on different platforms or settings. For example, the larger and more general *MFTC* data has a broad distribution of texts from all moral classes, but the *covid* dataset is heavily biased toward the *care/harm* foundation. People expressed more support and care when talking

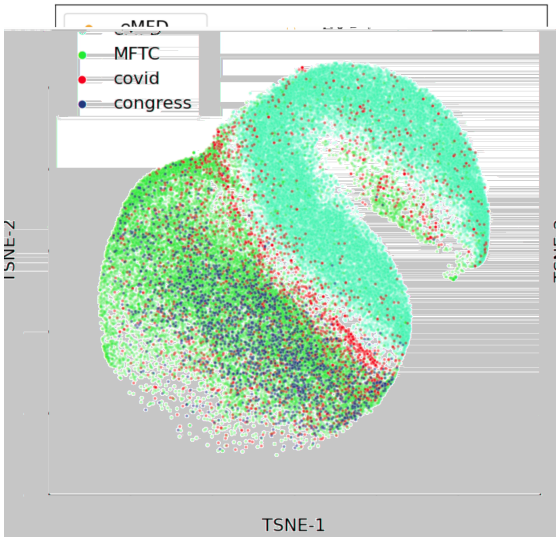

Figure 3: Feature distributions of different datasets, visualized with TSNE. For each text example (dots), we compute its embeddings using a vanilla BERT model without training, and then compute the first and second TSNE components.

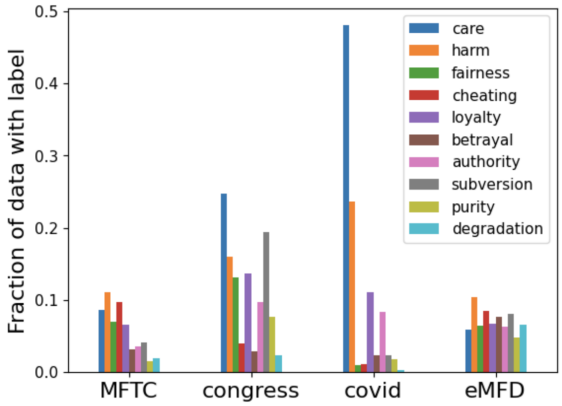

Figure 4: Label distributions of different datasets. The fraction of text examples with each label differs significantly among the datasets.

about pandemic-related topics such as "stay at home" (Fig. 2). This imbalance among classes can fundamentally affect multi-domain learning. For example, a model trained on the *covid* dataset learns very little about fairness. When a test data such as *MFTC* has many protest-related posts expressing fairness (Fig. 2 and Fig. 4), the model doesn't predict this label accurately (Table 2 row 1, BERT model performance).

Outstanding Performance of DAMF

We demonstrate how fusing these heterogeneous datasets with **DAMF** helps to tackle both feature and label shifts, and improves the model performance and generalizability. Table 2 shows the outstanding performance of **DAMF**. In most of the settings, it outperforms other baselines.

First, we compare **DAMF** with the widely used unsupervised method, DDR. It does not require training and thus is less susceptible to the training/test data shift problem. From Table 2 row 1 to 4, we see that DDR is not affected when training data are small and limited, which is an advantage when predicting on the larger general test data *MFTC*. On the other hand, the supervised methods BERT and **DAMF** have more difficulty when they are trained on smaller datasets. Nevertheless, as more data with annotations becomes available, supervised models gain more benefits. By merging these datasets and using a good model to improve generalizability, the full **DAMF** model outperforms DDR on all datasets.

Next, we compare **DAMF** with the current state-of-the-art morality prediction model, BERT. It is widely agreed that simply finetuning BERT can lead to overfitting on the training data and worsen its performance. Our results have shown that **DAMF** with domain adaptation ability fully outperforms the vanilla BERT, especially when the training and test data have more distinct feature and/or label distributions. One example is when training on *eMFD* and testing on *congress* (Table 2 row 9), where these two datasets' feature and label distributions are both distinct, **DAMF** outperforms BERT by 135% in F1-score.

We then discuss how **DAMF** compares to the strongest baseline, BERT+AFLite. On top of vanilla BERT, this method incorporates domain adaptation technique, filtering out overly easy training data points which may contain spurious correlation, and thus improves out-of-domain prediction performance. However, our results show that this method is not robust for all datasets. The *covid* dataset has a label distribution very concentrated on care and harm, and thus is easier to predict. In this case, BERT+AFLite has a good performance comparable to **DAMF** (Table 2 row 15, 17, 19 and 20). Nevertheless, when applied on *eMFD* with noisier texts and label distributions, AFLite even hurts the performance (e.g. BERT+AFLite has worse F1 than BERT in Table 2 row 6 and 11), possibly because AFLite mistakenly filters out relevant but noisy data.

In addition, we also observe from Table 2 that combining datasets for training generally improves BERT, BERT+AFLite and **DAMF**'s performances. Nevertheless, it is important to consider what datasets to merge. For example, *eMFD* has a more distinct feature distribution from others. When we merge *eMFD* and other dataset for training, it doesn't always improve the performance (e.g. comparing Table 2 row 5 and 7, row 12 and 14). By merging the suitable datasets with different feature and label distributions, we can facilitate the out-of-domain prediction by having the training data cover a wider feature distribution, and also by compensating each other to mitigate the label imbalance.

These above results show **DAMF**'s ability to handle feature and label shifts and to better adapt when learning multiple heterogeneous datasets. Table 3 shows some examples where **DAMF** successfully detected the correct moral foundations when other models fail. Row 3 and 4 show that BERT+AFLite does not always give robust predic-

	Training Data	Test Data	DDR	BERT	BERT+AFLite	DAMF -weightedL	DAMF full model
1	covid	MFTC	**0.38**	0.20 ± 0.01	0.20 ± 0.02	0.23 ± 0.04	0.26 ± 0.03
2	congress	MFTC	**0.38**	0.31 ± 0.02	0.30 ± 0.01	0.33 ± 0.01	0.35 ± 0.02
3	eMFD	MFTC	**0.38**	0.22 ± 0.03	0.20 ± 0.02	0.30 ± 0.04	0.34 ± 0.02
4	covid + eMFD	MFTC	**0.38**	0.24 ± 0.02	0.28 ± 0.02	0.27 ± 0.03	0.34 ± 0.02
5	covid + congress	MFTC	0.38	0.35 ± 0.01	0.34 ± 0.02	0.38 ± 0.01	$\mathbf{0.42 \pm 0.01}$
6	eMFD + congress	MFTC	0.38	0.34 ± 0.02	0.29 ± 0.02	0.36 ± 0.02	$\mathbf{0.39 \pm 0.02}$
7	covid + congress + eMFD	MFTC	0.38	0.31 ± 0.01	0.31 ± 0.02	0.33 ± 0.03	$\mathbf{0.43 \pm 0.01}$
8	covid	congress	0.23	0.10 ± 0.05	0.26 ± 0.01	0.12 ± 0.02	$\mathbf{0.35 \pm 0.02}$
9	eMFD	congress	0.23	0.14 ± 0.02	0.13 ± 0.02	0.21 ± 0.05	$\mathbf{0.33 \pm 0.01}$
10	MFTC	congress	0.23	0.30 ± 0.02	0.35 ± 0.01	0.30 ± 0.03	$\mathbf{0.38 \pm 0.01}$
11	covid + eMFD	congress	0.23	0.26 ± 0.03	0.21 ± 0.06	0.28 ± 0.02	$\mathbf{0.35 \pm 0.01}$
12	covid + MFTC	congress	0.23	0.31 ± 0.02	0.35 ± 0.01	0.34 ± 0.02	$\mathbf{0.40 \pm 0.03}$
13	eMFD + MFTC	congress	0.23	0.28 ± 0.05	0.34 ± 0.01	0.30 ± 0.03	$\mathbf{0.37 \pm 0.02}$
14	covid + eMFD + MFTC	congress	0.23	0.30 ± 0.01	0.32 ± 0.01	0.33 ± 0.02	$\mathbf{0.38 \pm 0.02}$
15	congress	covid	0.43	0.57 ± 0.01	$\mathbf{0.59 \pm 0.02}$	0.58 ± 0.01	$\mathbf{0.59 \pm 0.02}$
16	eMFD	covid	0.43	0.32 ± 0.03	0.23 ± 0.12	0.40 ± 0.05	$\mathbf{0.51 \pm 0.02}$
17	MFTC	covid	0.43	0.49 ± 0.02	$\mathbf{0.61 \pm 0.01}$	0.52 ± 0.01	0.56 ± 0.03
18	congress + eMFD	covid	0.43	0.47 ± 0.02	0.30 ± 0.02	0.51 ± 0.02	$\mathbf{0.57 \pm 0.01}$
19	congress + MFTC	covid	0.43	0.51 ± 0.01	$\mathbf{0.61 \pm 0.01}$	0.56 ± 0.02	$\mathbf{0.61 \pm 0.03}$
20	eMFD + MFTC	covid	0.43	0.45 ± 0.04	$\mathbf{0.57 \pm 0.03}$	0.47 ± 0.05	0.56 ± 0.02
21	congress + eMFD + MFTC	covid	0.43	0.49 ± 0.01	0.58 ± 0.02	0.52 ± 0.01	$\mathbf{0.61 \pm 0.02}$
22	covid	eMFD	0.17	0.12 ± 0.01	0.13 ± 0.01	0.13 ± 0.02	$\mathbf{0.18 \pm 0.01}$
23	congress	eMFD	0.17	0.16 ± 0.01	0.17 ± 0.01	0.17 ± 0.01	$\mathbf{0.20 \pm 0.01}$
24	MFTC	eMFD	0.17	0.13 ± 0.01	0.17 ± 0.01	0.16 ± 0.01	$\mathbf{0.18 \pm 0.01}$
25	covid + congress	eMFD	0.17	0.14 ± 0.01	0.16 ± 0.01	0.15 ± 0.01	$\mathbf{0.20 \pm 0.01}$
26	covid + MFTC	eMFD	0.17	0.14 ± 0.01	0.17 ± 0.01	0.15 ± 0.01	$\mathbf{0.19 \pm 0.01}$
27	congress + MFTC	eMFD	0.17	0.14 ± 0.01	0.17 ± 0.01	0.16 ± 0.01	$\mathbf{0.19 \pm 0.01}$
28	covid + congress + MFTC	eMFD	0.17	0.15 ± 0.01	0.17 ± 0.01	0.16 ± 0.01	$\mathbf{0.21 \pm 0.01}$

Table 2: Comparison between baseline models and **DAMF**. The numbers are F1 scores weighted by the number of true instances for each class. Each experiment is repeated with 5 random seeds to calculate the standard deviation. DAMF outperforms the baselines in most of the scenarios. The ablation study shows that both the domain adversarial module and the weighted loss function L^{MF} in moral foundation classifier play important roles.

tions. By outperforming strong baselines like BERT and BERT+AFLite in most experiments, **DAMF** is shown to achieve state-of-the-art performance for morality detection in out-of-domain testing.

Ablation Study

To verify that both the domain adversarial module and the weighted loss function help improve model generalizability, we perform an ablation study (the last two columns in Table 2). The DAMF-weightedL model uses the domain adversarial module and only a regular binary cross-entropy loss function. Figure 5 shows an example where the DAMF-weightedL model successfully aligns the distinct feature distributions between *eMFD* and *covid*. Furthermore, from Table 2 we see DAMF-weightedL already outperforms BERT model, showing the benefit of using the domain adversarial module. We also notice that the domain adversarial module is more effective when adapting between datasets with very distinct feature distributions (e.g. from *eMFD* to *congress* in row 9, DAMF-weightedL model outperforms BERT by 50% in F1). The domain adversarial module also tends to be more effective when adapting from smaller limited training datasets to larger general datasets (e.g. from *covid* to *MFTC* in row 1, DAMF-weightedL model outperforms BERT by 15% in F1). It is less effective when adapting from an already general training dataset to a limited test dataset, as the test data embeddings might already align with training data embeddings.

Next, we look at the performance of the full **DAMF** model which includes both the domain adversarial module and the weighted loss function. In all cases, the full model outperforms the DAMF-weightedL model. This says that the weighted loss function also plays an important role. This is especially true when the training data is *covid*. *covid* has the most imbalanced label distribution that is different from other datasets' label distributions, thus the weighted loss function gives a boost in performance. For example, when adapting from *covid* to *congress* (Table 2 row 8), the full **DAMF** outperforms the DAMF-weightedL model by 190%, and Table 4 shows the improved and more balanced per-class F1-scores by **DAMF**.

		predictions			
text	**ground truth**	**DDR**	**BERT**	**BERT+AFLite**	**DAMF**
Do u know how selfish you're being to put other peoples lifes at risk too????	betrayal	degradation	care	care	betrayal,harm
doing what they are supposed to be doing #socialdistancing	authority	fairness	subversion	subversion	authority
Stay home stay safe	care	care	non-moral	fairness	care
stay safe and stay home	care	care	non-moral	cheating	care

Table 3: Examples of **DAMF** successfully detected the correct moral foundations when other models fail. Row 3 and 4 also show that BERT+AFLite does not give robust predictions.

		DAMF-weightedL			DAMF		
	support	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
care	100	0.50	0.54	0.52	0.39	0.69	0.50
harm	62	0.47	0.42	0.44	0.35	0.56	0.43
fairness	45	0.00	0.00	0.00	0.06	0.11	0.08
cheating	17	0.00	0.00	0.00	0.09	0.41	0.15
loyalty	51	0.30	0.53	0.38	0.20	0.90	0.33
betrayal	12	0.00	0.00	0.00	0.06	0.50	0.11
authority	36	0.13	0.19	0.16	0.12	0.97	0.21
subversion	73	0.00	0.00	0.00	0.29	0.68	0.41
purity	31	0.00	0.00	0.00	0.13	0.97	0.22
degradation	10	0.00	0.00	0.00	0.03	0.70	0.06

Table 4: A comparison of DAMF-weightedL and **DAMF** on per-class performance. **DAMF** with the weighted loss has a more balanced performance among different label classes. Both models are trained on *covid* and tested on a sample of *congress* of size 370.

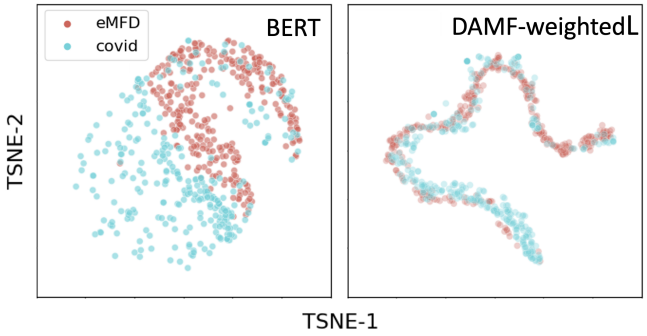

Figure 5: The feature embedding distributions of random samples of *eMFD* and *covid* datasets learned by vanilla BERT (left) and DAMF-weightedL (right), visualized with TSNE. DAMF-weightedL successfully aligns the feature distribution of these two distinct datasets.

Conclusions

We have proposed a data fusion framework, **DAMF**, for training moral foundation classifiers on heterogeneous datasets. We demonstrate the benefit of merging suitable datasets and show that **DAMF** outperforms three different baselines in various experimental settings. We also show that the domain adversarial module and the weighted loss function help align the distributions of data in the feature space and the label space, respectively.

There are limitations in our work that requires future improvements. By merging heterogeneous datasets using **DAMF**, we hope to reduce the differences between datasets due to varying procedures for data collection, annotation, topic selection, etc. One limitation is that we do not consider the complexities of various topics existing in single datasets like *MFTC*. One of our future works is to disaggregate the large datasets into different topics, and apply domain adaptation methods on separate topics. In addition, we could even disaggregate datasets by considering each annotator as a single domain, and target the differences from the annotation processes (Zhang et al. 2021).

To mitigate label shift between training and test data, the weighted loss function is a simple method. In the future, we plan to explore more sophisticated methods for domain adaptation targeting label shift, for example, estimating the relative class weights between domains and sample re-weighting as in the work of Tachet et al. (2020), or using additional loss functions to account for conflicting and co-occurring labels to achieve better generalization performance with respect to label shift (Kim et al. 2022).

Appendix

In Table 5, we list all the hyperparameters used fro training **DAMF** models under different experiment settings. We perform grid search for λ^{rec} in $[0, 0.1, 0.5, 1]$, λ^{trans} in $[no\ trans, 0.01, 0.1, 1, 10]$ and γ in $[0.1, 1, 10]$.

	Training Data	Test Data	λ^{trans}	λ^{rec}	γ
1	covid	MFTC	0.01	1	1
2	congress	MFTC	0.1	0.1	10
3	eMFD	MFTC	no trans	0	0.1
4	covid + eMFD	MFTC	1	0	0.1
5	covid + congress	MFTC	1	1	1
6	eMFD + congress	MFTC	0.1	1	10
7	covid + congress + eMFD	MFTC	no trans	0	0.1
8	covid	congress	0.1	0.1	1
9	eMFD	congress	10	0.1	1
10	MFTC	congress	no trans	1	1
11	covid + eMFD	congress	0.1	0.1	1
12	covid + MFTC	congress	no trans	0.1	1
13	eMFD + MFTC	congress	no trans	0	0.1
14	covid + eMFD + MFTC	congress	no trans	0	0.1
15	congress	covid	10	0.5	10
16	eMFD	covid	0.1	0.5	1
17	MFTC	covid	no trans	0	0.1
18	congress + eMFD	covid	0.01	1	10
19	congress + MFTC	covid	10	0	0.1
20	eMFD + MFTC	covid	no trans	0.1	1
21	congress + eMFD + MFTC	covid	no trans	0	0.1
22	covid	eMFD	0.01	0.1	1
23	congress	eMFD	0.01	0	1
24	MFTC	eMFD	no trans	0	0.1
25	covid + congress	eMFD	0.1	1	1
26	covid + MFTC	eMFD	no trans	0	1
27	congress + MFTC	eMFD	no trans	0	0.1
28	covid + congress + MFTC	eMFD	0.1	0	0.1

Table 5: Hyperparameters used for DAMF-weightedL model and the full DAMF model to produce results shown in Table 2.

Ethical Statement

Morality is relatively personal and subjective. It can vary based on different individuals, backgrounds or cultures. In this study, we focus on English language, American culture, and the morality expressed in news articles or on social media platforms. Unfortunately, in the datasets or the pretrained models that we have used, it is possible that biases exist with respect to gender, race, or other factors.

Acknowledgements

This project was funded in part by DARPA under contract HR001121C0168.

References

Alipourfard, N.; Burghardt, K.; and Lerman, K. 2021. Disaggregation via Gaussian regression for robust analysis of heterogeneous data. In *Handbook of Computational Social Science, Volume 2*, 269–288. Routledge.

Allaway, E.; Srikanth, M.; and McKeown, K. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 120–128. Sydney, Australia: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 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)*, 653–670. Online: Association for Computational Linguistics.

Fulgoni, D.; Carpenter, J.; Ungar, L.; and Preoţiuc-Pietro, D. 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)*, 3730–3736.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016.

Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Garten, J.; Boghrati, R.; Hoover, J.; Johnson, K. M.; and Dehghani, M. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Garten, J.; Hoover, J.; Johnson, K.; Boghrati, R.; Iskiwitch, C.; and Dehghani, M. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods, Instruments, and Computers*, 50(1): 344–361.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Graham, J.; Nosek, B. A.; Haidt, J.; Iyer, R.; Koleva, S.; and Ditto, P. H. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2): 366.

Guo, J.; Shah, D.; and Barzilay, R. 2018. Multi-Source Domain Adaptation with Mixture of Experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4694–4703. Brussels, Belgium: Association for Computational Linguistics.

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

Haidt, J.; and Joseph, C. 2004. Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, 133(4): 55–66.

Haidt, J.; Joseph, C.; et al. 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind*, 3: 367–391.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.

Hofmann, V.; Dong, X.; Pierrehumbert, J.; and Schuetze, H. 2022. Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 536–550. Seattle, United States: Association for Computational Linguistics.

Hofmann, W.; Wisneski, D. C.; Brandt, M. J.; and Skitka, L. J. 2014. Morality in everyday life. *Science*, 345(6202): 1340–1343.

Hoover, J.; Portillo-Wightman, G.; Yeh, L.; Havaldar, S.; Davani, A. M.; Lin, Y.; Kennedy, B.; Atari, M.; Kamel, Z.; Mendlen, M.; Moreno, G.; Park, C.; Chang, T. E.; Chin, J.; Leong, C.; Leung, J. Y.; Mirinjian, A.; and Dehghani, M. 2020. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8): 1057–1071.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 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(1): 232–246.

Islam, T.; and Goldwasser, D. 2022. Understanding COVID-19 Vaccine Campaign on Facebook using Minimal Supervision. *arXiv preprint arXiv:2210.10031*.

Johnson, K.; and Goldwasser, D. 2018. Classification of Moral Foundations in Microblog Political Discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 720–730. Melbourne, Australia: Association for Computational Linguistics.

Kennedy, B.; Atari, M.; Davani, A. M.; Hoover, J.; Omrani, A.; Graham, J.; and Dehghani, M. 2021. Moral concerns are differentially observable in language. *Cognition*, 212: 104696.

Kim, D.; Tsai, Y.-H.; Suh, Y.; Faraki, M.; Garg, S.; Chandraker, M.; and Han, B. 2022. Learning Semantic Segmentation from Multiple Datasets with Label Shifts. *arXiv preprint arXiv:2202.14030*.

Le Bras, R.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M.; Sabharwal, A.; and Choi, Y. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, 1078–1088. PMLR.

Li, Y.; Zhao, C.; and Caragea, C. 2021. Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6332–6345. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lin, Y.; Hoover, J.; Portillo-Wightman, G.; Park, C.; Dehghani, M.; and Ji, H. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, 552–559. IEEE.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer Feature Learning with Joint Distribution Adaptation. In *2013 IEEE International Conference on Computer Vision*, 2200–2207.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics*, 206–219. Springer.

Mooijman, M.; Hoover, J.; Lin, Y.; Ji, H.; and Dehghani, M. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6): 389–396.

Pacheco, M. L.; Islam, T.; Mahajan, M.; Shor, A.; Yin, M.; Ungar, L.; and Goldwasser, D. 2022. A Holistic Framework for Analyzing the COVID-19 Vaccine Debate. In *Proceedings of the 2022 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, 5821–5839. Seattle, United States: Association for Computational Linguistics.

Rojecki, A.; Zheleva, E.; and Levine, L. 2021. The Moral Imperatives of Self-Quarantining. *Annual meeting of the American Political Science Association*.

Roy, S.; and Goldwasser, D. 2021. Analysis of Nuanced Stances and Sentiment Towards Entities of US Politicians through the Lens of Moral Foundation Theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 1–13. Online: Association for Computational Linguistics.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3): 258–268.

Tachet des Combes, R.; Zhao, H.; Wang, Y.-X.; and Gordon, G. J. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33: 19276–19289.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazazian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; et al. 2022. The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*.

Wang, S.-Y. N.; and Inbar, Y. 2021. Moral-language use by US political elites. *Psychological Science*, 32(1): 14–26.

Xie, J. Y.; Hirst, G.; and Xu, Y. 2020. Contextualized moral inference. arXiv:2008.10762.

Zhang, X.; Xu, G.; Sun, Y.; Zhang, M.; and Xie, P. 2021. Crowdsourcing learning as domain adaptation: A case study on named entity recognition. *arXiv preprint arXiv:2105.14980*.

Zhang, Y.; Barzilay, R.; and Jaakkola, T. 2017. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 5: 515–528.