Estimating the transferability of state-of-the-art models in predicting moral values

Alin Dondera[1] , **Enrico Liscio**[1] , **Pradeep Murukannaiah**[1]
[1]TU Delft

Abstract

Moral values play a crucial role in our decision-making process by defining what is right and wrong. With the emergence of political activism and moral discourse on social media, and the latest developments in Natural Language Processing, we are looking at an opportunity to analyze moral values to observe trends as they form. Recent studies have extensively examined the performance of different NLP models for estimating moral values from text, but none of them has tackled the problem of transfer learning. Our study provides a comprehensive look into the cross-domain performance of three state-of-the-art models. We find that BERT, the current most used model in Natural Language Processing, offers the best results. For reproducibility, we publicly release our code on GitHub.

Keywords: Moral values, moral foundation theory, transfer learning, domain adaptation, natural language processing

1 Introduction

Moral values are the abstract motivations that drive our opinions and actions. Understanding these values can, in turn, help researchers understand what divides people and what can be done to overcome these divides. In the context of Artificial Intelligence (AI), understanding moral values is essential in achieving beneficial AI, aimed at the creation of value-aligned artificial agents that can operate among us [1]. However, as Graham et al. claim, “morality varies greatly across cultures, as do individuals within cultures” [2]. Due to the subjective nature of values, it is difficult for AI systems to accurately predict these values [3] [4].

Moral Foundations Theory (MFT) [5] proposes that morality can be expressed in terms of five moral foundations. MFT is one of the most established theories in the field of social sciences due to its practicality and openness to new changes. Alongside MFT, the Moral Foundation Questionnaire [6] was developed to aid researchers in collecting information about the morality of individuals. These standardized questionnaires, however, are limited by the “artificial nature of the stimuli used and the non-natural settings in which they are embedded” [7].

A more natural and scalable environment to estimate moral values can be found in social media. Mediums such as Facebook and Twitter are used every day by people and can therefore better portray what moral values are present in a person’s natural environment. This, however, requires the use of Natural Language Processing (NLP) models to extract meaningful information from textual content.

Recently, MFT was applied in the field of computational sciences with the development of the Moral Foundation Twitter Corpus (MFTC) [8]. This dataset contains 35 thousand tweets, split into seven sub-datasets, annotated with the MFT schema for their moral values. The presence of the MFTC dataset allows an in-depth analysis of the performance of state-of-the-art NLP models (such as BERT) [9] [10] in predicting moral values. By evaluating the performance of these NLP models, we can demonstrate a potential use case for real-world applications (e.g. using anonymous Twitter data to observe trends in moral values as they form [3]).

In order to assess the usability of our analysed models in real-world scenarios, we need to take a look at the transferability of our models. Transferability is the extent to which a machine learning model can “transfer knowledge learned in one or more source tasks and use it to improve learning in a related target task” [11]. In this paper, we are interested in evaluating the transferability of our models across different domains (domain adaptation). Knowing that, according to Moral Foundation Theory, moral values transfer across domains, we want to know whether the representations learned by our models do as well.

An in-depth understanding of the performance of domain adaptation could imply that it is possible to pre-train bigger models on larger datasets for predicting moral values that can later be fine-tuned. These models can then be made available to the broader community through open-sourcing, without the need for massive computation resources or data. This could be especially useful for social scientists who can make use of such models to analyze social media trends, such as predicting the emergence of violence during protests using Twitter data [3].

2 Related Work

Moral values play an essential role in AI development, as they can ensure that powerful AI systems are properly aligned with human values. This fact becomes increasingly important as the autonomy and speed of these systems increase [12]. To better understand morality, Graham and Haidt introduced the Moral Foundation Theory (MFT) [5], which expresses morality in terms of five moral foundations, with each foundation being comprised of a vice and a virtue.

In recent years, numerous studies that combined the fields of NLP and moral values (operationalized by MFT) have been conducted. Most of these studies rely on the Moral Foundation Dictionary (MFD) [13], a manually created vocabulary of words, where each word is associated with a set of moral values. This approach uses the dictionary to calculate the frequency of moral values in specific texts and uses that as input for machine learning models.

In 2014, Dehghani et al. [14] applied MFD and topic modelling (Latent Dirichlet Association) to analyze conservative and liberal weblogs and found differences between the moral frameworks of the two groups. Two years later, Teernstra et al. [15] applied machine learning models for classifying moral values in tweets related to the political debate surrounding ‘Grexit’. Their study advocates for the use of a more *pure* machine learning approach in contrast to the dictionary approach. Following these studies, Lin et al. [16] adopted a new method that acquires background knowledge from Wikipedia to enrich the information captured in their input vectors.

Recently, Hoover et al. [8] published the Moral Foundation Twitter Corpus (MFTC), a dataset of 35 thousand tweets, annotated for moral sentiment using the MFT schema. The MFTC is divided into seven sub-datasets, each related to an influential trend on social media. Alongside the dataset, Hoover et al. also report a set of classification baselines for different machine learning approaches. Their results suggest that recurrent neural networks outperform dictionary-based approaches.

The current state-of-the-art results in predicting moral values are achieved by Araque et al. [17], which extend the MFD to obtain a new lexicon entitled *MoralStrength*. The authors then apply this lexicon to the MFTC corpus, training a model for each combination of moral foundation and corpus sub-dataset. Each model is trained to predict whether a text belonging to a specific sub-dataset contains a specific moral foundation without differentiating between vice and virtue.

None of these studies approaches the problem of transfer learning, specifically the cross-domain performance of moral value prediction. Other than Lin et al. [16], which specifically mention that their models are not suited for migration to new domains, the rest of the studies do not mention the concept of transferability or generalizability. It is possible that the current state-of-the-art, the models trained using *MoralStrength*, may also suffer the same issue since, by training on specific domains, they are prone to overfitting. Our study aims to fill this research gap by evaluating the cross-domain performance of different models in predicting moral values.

3 Methodology

Our work is concerned with estimating the transferability across domains of Natural Language Processing (NLP) models that are trained on predicting moral values from social media text. Predicting moral values is a text classification task similar to sentiment analysis [18][19], with some key differences. Firstly, this is a multi-label classification problem, as opposed to a multi-class problem. This means that texts can be labeled with multiple moral values at the same time. Secondly, as mentioned in [16], moral values are closely related to their relevant context, hence they require a better understanding of background knowledge.

We can formally define our task as follows: given a set of texts $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ and a set of moral values $\mathcal{L} = \{l_1, l_2, \ldots, l_m\}$, we wish to learn a mapping $f : \mathcal{T} \mapsto \mathcal{P}(\mathcal{L})$. The elements of $\mathcal{P}(\mathcal{L})$ are represented as binary vectors of form $y = \{y_1, y_2, \ldots, y_m\}$, where $y_i = 1$ only if the text is labeled with the label l_i. The mapping f will be obtained by employing three commonly used NLP models: Recurrent Neural Networks (LSTM) [20], FastText [21], and Transformers (BERT) [9].

In our experiments, we will evaluate the transferability of each of these models by first partitioning the set of texts $\mathcal{T}$ into a source domain $\mathcal{T}_{source}$ and a target domain $\mathcal{T}_{target}$. Next, we will learn a mapping f on the source domain, also defined as the pre-training step, and use transfer learning techniques to transfer the knowledge of this mapping to the target domain, also defined as the fine-tuning step.

In the first stage, we distinguish between three transfer learning approaches that will be evaluated: *simple*, *train all* and *fine-tune*. The *simple* approach learns a mapping on the source domain and directly uses it for predictions on the target domain. This approach offers a good estimation of the model’s generalizability. Next, the *fine-tune* approach uses the mapping learned on the source domain, and updates it on the target domain. We expect this technique to perform best, but it may cause overfitting to this new domain, leading to poor performances on the source domain. Thus, we use the *train all* approach which combines the train sets of the source and target domain into a single train set, which is then used to learn a new mapping. We expect this method to obtain similar performance to the *fine-tune* approach, while overfitting less, at the expense of runtime. Lastly, we include a *no pre-train* approach, which learns the mapping directly on the target domain, without utilizing the source domain. This approach will serve as a baseline and will show if any of the aforementioned transfer learning approaches increase performance.

Consequently, in the second stage, we will enhance the cross-domain performance of the BERT model by applying state-of-the-art domain adaptation techniques. For this, we will utilize the solutions described in [22], namely further pre-training and layer-wise learning rates. For further pre-training, we will use a BERT model pre-trained on data with a similar distribution as our source and target domains. Lastly, we try to exploit the fact that the different layers of a neural network can capture different types of information [23], by setting a different learning rate for the parameters of each layer.

4 Experimental setup

Before evaluating the transferability of the models mentioned above, we need first to describe the dataset used in our experiments: the Moral Foundation Twitter Corpus. Furthermore, we will offer an overview of the general architecture of the studied models. Lastly, we will highlight some important implementation details needed to properly understand the results of our experiments.

4.1 Data

For our experiments, we have used the Moral Foundation Twitter Corpus [8] (MFTC) dataset, available online[1] for anyone to use. The dataset is composed of 35 thousand tweets, divided into seven separate sub-datasets, each corresponding to a specific topic, namely: All Lives Matter (ALM), Black Lives Matter (BLM), Baltimore protests, hurricane Sandy, the MeToo movement, the 2016 presidential election, and hate speech and offensive language [24]. This diverse array of datasets from complex socio-political issues makes it possible for us to evaluate the transferability across domains by treating each different sub-dataset as a unique domain.

The tweets were hand-annotated for their moral values by several annotators by making use of Moral Foundation Theory (MFT). MFT is a psychological theory that argues for the existence of five moral foundations, from which different moralities can be created. Each of these moral foundations is comprised of a virtue, representing what is ethically right, and a vice, representing what is ethically wrong. These foundations, along with their definitions, can also be seen in Table 1:

Foundation	Definition
Care Harm	This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.
Fairness Cheating	This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy
Loyalty Betrayal	This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
Authority Subversion	This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
Purity Degradation	This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

Table 1: Definitions of moral foundations. Taken from [25].

Due to the subjective nature of these moral values, different annotators may label the same tweet differently. In order to assign a unique target vector to each of these tweets, we applied a majority vote, similar to the original paper [8]. This means that a tweet was labeled with a specific moral value, if and only if at least half the annotators agreed on that value. Tweets where no such agreement was possible were labeled with a *non-moral* label. The frequency of tweets per label for each sub-dataset can be seen in Table 2.

From this distribution of labels, the imbalanced nature of our dataset becomes apparent. Some sub-datasets suffer significantly from this imbalance, such as Davidson and Baltimore, where the number of *non-moral* labels greatly outweigh the rest. It is important to quantify the imbalancedness, since it can be useful when interpreting results. For this, we use the methods described in [26]: the imbalance ratio per label (IRLbl), the mean imbalance ratio (MeanIR), and the coefficient of variation of IRLbl (CVIR). Each of these metrics is calculated relative to the majority label in the dataset (the *non-moral* label in our case). These metrics are calculated as follows:

- **IRLbl**: For each label y, it is calculated as the ratio between the majority label and label y.
- **MeanIR**: It is calculated as the mean of all IRLbl. This offers an overall idea of the imbalanced nature of the dataset, but it should be used alongside CVIR to get a clearer picture.
- **CVIR**: It is calculated as the ratio between the standard deviation of all IRLbl and the MeanIR. A high CVIR will indicate that the degree of imbalancedness differs greatly between labels, while a low CVIR indicates that all labels suffer from similar degrees of imbalancedness.

Table 3 shows these metrics applied to our sub-datasets.

Dataset	MeanIR	CVIR
ALM	11.54	1.10
Baltimore	51.26	1.40
BLM	5.35	0.77
Davidson	344.84	1.13
Election	9.62	0.67
MeToo	3.99	0.62
Sandy	6.37	1.10

Table 3: MeanIR and CVIR per sub-dataset.

4.2 Models

We have chosen three main machine learning models for evaluating their cross-domain transferability, namely: Long Short Term Memory neural networks (LSTM), FastText and Bidirectional Encoder Representations from Transformers (BERT).

LSTM

LSTMs, introduced in 1997 [20], have been widely used in NLP tasks, and have recently been employed in several studies for predicting moral values [3] [4] [8] [16]. They are part

[1] The dataset and instructions for how it can be used can be found at https://osf.io/k5n7y/

Moral Value	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
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
Non-moral	1744	3826	1583	4509	2501	1565	1313

Table 2: Distribution of labels per sub-dataset

of a specific category of deep learning models, called Recurrent Neural Networks (RNN). In contrast to feedforward neural networks, RNNs process input sequentially while keeping an internal state (memory).

LSTMs were considered state of the art for tasks such as sequence to sequence learning and machine translation in the past, but have fallen short to new models in recent years. The reason for this is that LSTMs (and RNNs in general) suffer from two major problems. Firstly, due to their sequential nature, they cannot make good use of GPUs, hence training is slower than for other networks. Secondly, they suffer from the exploding gradients problem during training [27]. Exploding gradients are large accumulated errors that cause substantial updates to the parameters, thus making the network unstable.

We will still include LSTMs in our evaluation, since they offer a good baseline for comparison with the other models. This choice is also motivated by their wide use in recent studies, as previously mentioned.

FastText

FastText is a machine learning library designed for learning word representations and text classification. This architecture attains scores on par with previous state-of-the-art deep learning methods, while being considerably faster [21]. Another advantage of this model is that it does not require a GPU to handle training, by solely relying on the CPU, making it more available. This is especially important in real-time applications.

FastText also learns character-level information, in contrast to LSTMs, which only focus on whole word representations [28]. This allows FastText to better deal with words that rarely appear or not appear at all in the vocabulary. In the context of transfer learning, it is also interesting to evaluate the transferability of the character-level information learned by the model.

BERT

Bidirectional Encoder Representations from Transformers (BERT) [9] is a language representation model based on the Transformer architecture [29]. The transformer architecture gave rise to several pre-trained language models, becoming one of the most popular approaches in NLP for transfer learning.

The model is comprised of an embedding layer, followed by a stack of 12 identical encoders. Before feeding the input to the model, the text has to be preprocessed. Each sentence is split into a list of tokens, to which a unique [CLS] token is inserted at the first position. The embedding layer then takes this list of tokens as input and transforms each of them into a fixed-size representation. The list of embeddings is then fed on to the first encoder. Each encoder will receive a list of embeddings and would, in turn, output a new list of embeddings of the same size. An illustration of this process can be seen in Figure 1.

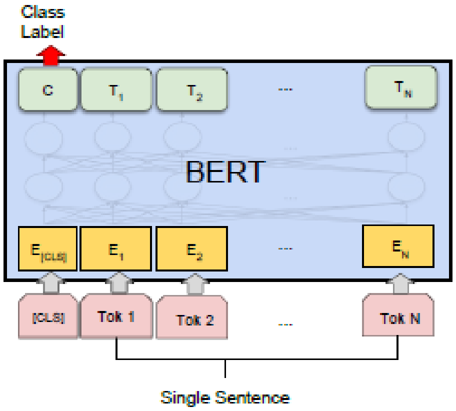

Figure 1: Overview of BERT architecture. Taken from [9]

For sentence classification tasks, such as predicting moral values, the authors of [9] suggest using the final embedding of the [CLS] token. It represents an ”aggregate sequence representation for classification tasks”. In order to predict a moral value based on this vector, we feed it through a simple feedforward neural network with an 11-dimensional output layer[2].

BERT has several advantages compared to the two previous models. Due to its architecture, it can learn contextual representations for words. This means that the same word may have different embeddings in different contexts. Moreover, since data flows in parallel through the model, BERT can benefit from parallelization, which significantly speeds up processing.

Currently, variants of BERT obtain the best performances across several NLP tasks, including sentiment analysis [30], which makes it a great candidate for our task. One reason BERT achieves these state-of-the-art results is its ability to transfer knowledge between similar domains and tasks. Therefore, it is essential to know for our case if this transferability across domains also applies to a more abstract concept, such as moral values.

4.3 Implementation details

For multi-label classification tasks, the F1 score is the most widely used metric. In our experiments, we will focus on the Macro F1 score. We do this because we want all classes to contribute equally to the end result. The dataset suffers from bias to specific labels, thus having each class contribute equally offers a more unbiased estimation.

In both stages, we will apply the same overall technique for obtaining the results. We will partition the dataset into two separate domains, A and B. Domain A will be called the source domain and will contain six out of the seven sub-datasets from MFTC. Domain B will be called the target domain and will contain the remaining sub-dataset. The source domain will be used for training a model, while the target domain will be used to evaluate the out-of-domain performance of that model. We will run our experiments seven times, with each sub-dataset from the MFTC corpus appearing once as the target domain.

A useful technique in assessing the performance of a model is cross-validation. For our experiments, we are using 10-fold cross-validation, similar to [8]. We start by splitting both the source domain and the target domain into 10 folds. At each iteration, we choose one of these folds as a holdout test set and use the other nine for training (or fine-tuning for the target domain). The final result is the average Macro F1 over these 10 runs. This method ensures that we end up using all data for testing, which reduces the variance of our results. To make experiments comparable, we made sure that they used the same data by setting random seeds.

Lastly, before feeding the text to the respective models, we applied a preprocessing step to the textual content of the tweets. This was done by normalizing commonly encountered syntax such as URLs, emails, usernames and mentions. Moreover, common spelling mistakes were corrected, and contractions were unpacked. This was done using the Ekphrasis package [31]. Emojis were also transformed to their respective words using the Python Emoji package.

5 Results and Discussion

The results were obtained using the parameters found in Appendix A. The environment used for running the experiments is the High-Performance Computing (HPC) cluster at TUDelft[3].

5.1 Comparing Models

In this experiment, we evaluated the cross-domain performance of our models (LSTM, FastText, and BERT) using the four aforementioned approaches: *no pre-train*, *simple*, *fine-tune*, and *train all*. The results are summarized in Table 4.

It is apparent from this table that the pre-training step offers a significant increase in performance. This is most clear in the case of LSTM, where after fine-tuning, we observe a 16% rise in the average F1 score, with fastText and BERT gaining a meaningful 9% increase as well.

Looking at Table 4, we also observe the expected relationship between the three transfer learning strategies, namely: *simple* $\prec$ *train all* $\prec$ *fine-tune*. This applies to all combinations of models and datasets, with the only exceptions surrounding the Davidson sub-dataset. Its highly imbalanced nature can explain this behaviour. For example, it is possible that models like LSTM and BERT, which make use of word-level information, may overfit to certain offensive words and label all samples as *non-moral*.

It is also clear from the results that, from the three models, BERT outperforms the rest when it comes to transferability. This was also the expected result, as BERT represents the current state-of-the-art in the field of NLP. FastText is the next best-performing model by a significant margin. While it manages to do so in a fraction of the time needed by BERT [32, Figure 10], we believe that runtime is not an influential metric in the context of transferability, since the pre-training step only needs to be done once. Both these models outclass the LSTM, which is the most extensively used in literature for predicting moral values (e.g [3] [4] [8] [16]).

Another important finding relates to BERT’s generalizability. We observe from our results that BERT, without further retraining on the target domain, obtains better results than a fine-tuned LSTM and is relatively close to the fine-tuned fastText, by a 2.8% margin. This result is impressive since, as mentioned in [8] and [16], predicting moral values relies heavily on understanding the target domain’s context.

Lastly, we observe a discrepancy between the scores of each target domain. This difference is most noticeable for two sub-datasets: Baltimore and Davidson. Looking at Table 3, we also see that these two datasets have a high MeanIR and a high CVIR, which can explain the low scores. On the other hand, the BLM sub-dataset obtains considerably higher scores on all counts compared to the others. However, we find no direct correlation between this and the metrics reported in

[2] For a more detailed description of BERT, please have a look at the original papers on transformers [29] and BERT [9]. There are also multiple online resources for this such as this Transformer tutorial or this BERT tutorial.

[3] See login.hpc.tudelft.nl

Table 3. The reason for this behaviour could be embedded in the annotation procedure. While annotating the MFTC, the BLM sub-dataset was largely annotated by five annotators, while the other sub-datasets were only annotated by three or four annotators [8, Table 1]. This difference in the number of annotators could result in less noise in the data and better predictions. Another possible explanation could be that the language used in the BLM sub-dataset lacks ambiguity, which would make our models able to better associate word representations with specific moral values.

5.2 Enhancing BERT

As a result of BERT's performance, we believed it is in the best interest of this research to further try and improve its performance while fine-tuning on the target domain, by using current state-of-the-art domain adaptation techniques. In doing so, we will explore the solutions described in [22]. The results of our findings can be found in Table 5.

We first extend our model by further pre-training BERT on similar data (English tweets). We employ the already-existing BERTweet [19] model, which was pre-trained on 850M English tweets + 23M tweets surrounding the COVID-19 pandemic. When applied to our dataset, it results in a slight increase in performance (0.5%).

Secondly, we extend our fine-tuning implementation to use a layer-wise decreasing learning rate (LLR). We do this by setting a learning rate for the last layer of the BERT model and subsequently multiplying it with a decay factor for each lower layer. Implementing this method provided us with a 0.7% increase in F1 score compared to the previous approach.

6 Responsible Research

Estimating moral values is a complex task with several ethical ramifications. As researchers, we have a responsibility to both the research community and to the people that may end up being affected (either positively or negatively) by the outcome of this paper.

In this section, we start off by highlighting the steps we took to ensure the reproducibility of our results. Next, we will try to cover the ethical implications of our software on society in general, both positive and negative.

Reproducibility

Reproducibility is an essential concept in the field of science due to multiple reasons. First, it is a good indicator of the credibility of the research that was carried. Without this credibility, the results cannot be considered scientific knowledge. Second, reproducible research also has the advantage of obtaining possible validation of results from other parties who wish to build upon the results. Moreover, reproducible results can also aid these parties in extending and improving the results or in gaining new insights by offering them a solid base to build from.

Research into artificial intelligence, and deep learning, in particular, is more difficult to reproduce due to the massive amount of parameters and randomness (e.g. training kickoff) these deep learning models depend on. Due to the complicated nature of some of our models, we have to deal with these same issues. To ensure reproducibility to the best extent possible, we have inspired from the steps outlined in [33].

Most importantly, we have made the code available on the project's GitHub page. The code contains the same scripts that were used for running the experiments. In addition, the random seeds used internally by the libraries are fixed from the beginning to ensure that the results will remain the same between different runs.

The data used for running the experiments can be obtained by contacting the authors of the Moral Foundation Twitter Corpus [8]. If the authors cannot supply the data, the texts can also be retrieved using the Twitter API. It is important to note that some of these tweets may have been removed due to violating platform rules (e.g., messages containing hate speech, racism). This difference in data could lead to slightly different, however, still comparable results.

Ethical Considerations

With the increasing effects that engineering innovations have on everyday life, as members of the engineering community, it is crucial for us to think about how our research may affect people. Once introduced into society, the effects of a particular technology are difficult to stop. Hence it is essential to think about these consequences from the earlier design stages. It is important to note that the consequences are still diffuse across time and geographical location (e.g., the software can have different consequences years from now, or it can have different consequences in different parts of the world), hence an exhaustive list of implications is not feasible.

With that said, we will first highlight the possible positive impacts of our research. The goal of our research is to estimate the transferability of machine learning models in predicting moral values. We are studying this because transfer learning allows the engineering community to use complex models more efficiently. Since our results suggest that these models are indeed transferable, it means that we can make these models available to social scientists who better know how to use them. We showed that better results could be achieved with the current state of the art, hoping that these results can improve the results of other studies in turn. These improved results can then be used to, for example, design value-aligned AI or to analyze data in real-time and observe the formations of new trends or predict the emergence of violence during protests, as previously done by [3].

However, there are some negative implications to our research as well. Firstly, our models deal with social media (Twitter) data, which could lead to privacy issues if used in real-world applications. Different persons may feel like their privacy is invaded if their words are analyzed for morality. This issue can be tackled through consent forms, for example, where people can explicitly refuse the collection of their data for this purpose. Since predicting moral values has no direct benefits for them, users may not give consent, leading to a scarcity of data. Nevertheless, our results in transferability may prove helpful in this regard since transfer learning reduces the need for data.

Secondly, there has been growing concern regarding the environmental impact of deep learning models [34]. Since BERT, our best-performing model, has hundreds of millions

Experiment	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy	Avg
LSTM No Pre-train	25.66	8.31	56.00	8.74	24.63	18.38	27.56	24.18
LSTM Simple	40.27	20.06	43.94	9.11	36.68	32.70	22.01	29.25
LSTM Train All	48.28	22.13	63.36	9.14	41.62	41.12	34.51	37.17
LSTM Fine-Tune	51.74	23.08	74.49	8.93	43.86	43.35	37.19	40.38
fastText No Pre-train	52.36	18.17	74.11	8.73	39.39	40.88	34.03	38.24
fastText Simple	47.18	22.96	54.27	9.21	39.85	44.72	23.98	34.60
fastText Train All	50.90	24.79	58.41	9.47	44.74	48.26	45.18	40.25
fastText Fine-Tune	57.41	29.75	78.95	9.74	52.10	59.46	47.11	47.79
BERT No Pre-train	58.13	24.28	78.95	8.74	53.16	55.17	43.69	46.02
BERT Simple	58.10	28.65	75.12	9.37	54.55	51.35	37.79	45.00
BERT Train All	66.78	36.29	84.85	9.46	62.65	59.58	55.13	53.53
BERT Fine-Tune	67.35	37.56	86.03	8.73	65.45	59.73	57.23	54.58

Table 4: Results for evaluating cross-domain transferability. The columns showcase the target domain on which the models were evaluated.

Experiment	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy	Avg
BERTweet	68.40	39.45	85.95	8.74	65.70	60.52	57.11	55.12
BERTweet + LLR	67.61	39.70	86.26	8.74	66.96	61.73	59.79	55.83

Table 5: Results of domain adaptation techniques applied to enhance BERT's transferability. LLR refers here to layer-wise learning rates.

of parameters, this effect cannot be overlooked. Similar to the previous point, transferability may be helpful in this regard as well. With transfer learning, by pre-training a model on a large corpus before, people can reuse this model while skipping an expensive training phase, which saves computation resources.

7 Conclusions and Future Work

Moral values are the factors that guide us in our decision-making process. By estimating these values from online discourse, we can observe trends as they form and understand what divides people and how to overcome these divides. However, while numerous studies have applied NLP techniques for predicting moral values in social media text (Tweets), no attention was paid to the transferability of the proposed models to unseen domains. This issue is crucial to consider as it directly relates to the usability of the models. A transferable model has several advantages: it requires less data, less runtime, and frequently results in better performance.

Our core contribution fills this gap by offering a comprehensive evaluation of the cross-domain performance of state-of-the-art NLP models in predicting moral values. Specifically, we evaluated three models (LSTM, fastText and BERT) by first pre-training them on a source domain, containing tweets from several topics, and then using this knowledge to evaluate tweets on an unseen target domain.

Our study shows that all three models can transfer knowledge to unseen domains. This can especially be seen in how they all benefit significantly from the pre-training step. From all three, BERT has the best results, followed by fastText and then LSTM. It is important to note that BERT obtains reasonably high scores without any need for retraining on new domains, being only 2% worse than the best performing fastText approach. This result suggests that BERT can generalize well to unseen domains, which is a remarkable result given how dependent moral values are on their underlying context.

Overall, BERT offers a 7% increase in performance compared to the next best performing model. We then enhanced this BERT model, by utilizing novel domain adaptation techniques, resulting in a further 1.2% gain in F1 score. However, we believe that there is still room for improvement for our results, either through better hyperparameter tuning or through different domain adaptation techniques. For example, the regularization approach outlined in [35] could be used to enhance the performance of our models, since it was able to attain state-of-the-art results in text classification tasks.

Lastly, we note that it is easier to transfer knowledge to some domains than others. While the low performance on some is directly correlated to their imbalanced nature, further research is required to understand what are the important factors in a *good* target domain. On the other hand, it would also be interesting to see the important factors in a *good* source domain. For example, in our case, we believe that training on a biased dataset such as Davidson may decrease performance.

References

[1] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *Ai Magazine*, vol. 36, no. 4, pp. 105–114, 2015.

[2] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in experimental social psychology*, vol. 47, Elsevier, 2013, p. 58.

[3] M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani, "Moralization in social networks and the emergence of violence during protests," *Nature human behaviour*, vol. 2, no. 6, pp. 389–396, 2018.

[4] R. Rezapour, S. H. Shah, and J. Diesner, "Enhancing the measurement of social effects by capturing morality," in *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2019, pp. 35–45.

[5] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007.

[6] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto, "Mapping the moral domain.," *Journal of personality and social psychology*, vol. 101, no. 2, p. 366, 2011.

[7] W. Hofmann, D. C. Wisneski, M. J. Brandt, and L. J. Skitka, "Morality in everyday life," *Science*, vol. 345, no. 6202, pp. 1340–1343, 2014.

[8] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, *et al.*, "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[11] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[12] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020.

[13] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations.," *Journal of personality and social psychology*, vol. 96, no. 5, p. 1029, 2009.

[14] M. Dehghani, K. Sagae, S. Sachdeva, and J. Gratch, "Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque"," *Journal of Information Technology & Politics*, vol. 11, no. 1, pp. 1–14, 2014.

[15] L. Teernstra, P. van der Putten, L. Noordegraaf-Eelens, and F. Verbeek, "The morality machine: Tracking moral values in tweets," in *International Symposium on Intelligent Data Analysis*, Springer, 2016, pp. 26–37.

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

[17] O. Araque, L. Gatti, and K. Kalimeri, "Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction," *Knowledge-based systems*, vol. 191, p. 105 184, 2020.

[18] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170–179, 2014.

[19] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 9–14.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" In *China National Conference on Chinese Computational Linguistics*, Springer, 2019, pp. 194–206.

[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[24] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[25] MoralFoundations.org. (2021). "Moral foundation theory," [Online]. Available: https://moralfoundations.org/. (accessed: 27.06.2021).

[26] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, 2015.

[27] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, PMLR, 2013, pp. 1310–1318.

[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *arXiv preprint arXiv:1607.04606*, 2016.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

[30] Papers with Code. (2021). “Papers with code - sentiment analysis,” [Online]. Available: https://paperswithcode.com/task/sentiment-analysis. (accessed: 25.06.2021).

[31] C. Baziotis, N. Pelekis, and C. Doulkeridis, “Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 747–754.

[32] A. Geadau, E. Liscio, and P. Murukannaiah, “Performance analysis of the state-of-the-art nlp models for predicting moral values,” TU Delft Bachelor Thesis, TU Delft Repository, 2021.

[33] V. C. Stodden, “Reproducible research: Addressing the need for data and code sharing in computational science,” 2010.

[34] K. Martineau. (2020). “Shrinking deep learning’s carbon footprint,” [Online]. Available: https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807. (accessed: 27.06.2021).

[35] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, “Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization,” *arXiv preprint arXiv:1911.03437*, 2019.

A Training details

Firstly, we describe the computing environment and software used for reproducing our results. The experiments were run on the High-Performance Computing(HPC) Cluster at TUDelft. The exact environment is as follows:

- GPU: GeForce RTX 2080 Ti (only used for LSTM and BERT)
- PyTorch: 1.8.1 (BERT)
- TensorFlow: 2.4.1 (LSTM)
- FastText: 0.8.1
- Hugginface’s Transformers: 4.6.0 (BERT)
- CUDA: 11.2 (only used for LSTM and BERT)
- cuDNN: 8.1.1.33 (only used for LSTM and BERT)

In continuation, we showcase the parameters used for running the experiments in Tables 6, 7, and 8. If a parameter for any of the models cannot be found in these tables, the default value supplied by the framework was used.

The results in Section 5.2 were obtained by first changing the *model name* to *vinai/bertweet-covid19-base-uncased*. Secondly, we introduced a learning rate decay rate of 0.95, and increased the learning rate of the last layer to $5 * 10^{-5}$.

Parameter name	Value
Word Embeddings	glove.6b.300d
Epochs	10
Batch size	128
Maximum sequence length	100
Optimizer	Adam
Learning rate	0.01
Loss function	Binary Cross Entropy

Table 6: Parameters used for running experiments with LSTM

Parameter name	Value
Epochs	50
Learning rate	0.03
Threshold	0.3

Table 7: Parameters used for running experiments with fastText

Parameter name	Value
Model name	bert-base-uncased
Epochs	3
Batch size	16
Maximum sequence length	64
Optimizer	AdamW
Learning rate	$5 * 10^{-5}$
Loss function	Binary Cross Entropy

Table 8: Parameters used for running experiments with BERT