Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Measuring Moral Dimensions in Social Media with Mformer

Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, Lexing Xie

Australian National University
{josh.nguyen, ziyu.chen, nicholas.carroll, alasdair.tran, colin.klein, lexing.xie}@anu.edu.au

Abstract

The ever-growing textual records of contemporary social issues, often discussed online with moral rhetoric, present both an opportunity and a challenge for studying how moral concerns are debated in real life. Moral foundations theory is a taxonomy of intuitions widely used in data-driven analyses of online content, but current computational tools to detect moral foundations suffer from the incompleteness and fragility of their lexicons and from poor generalization across data domains. In this paper, we fine-tune a large language model to measure moral foundations in text based on datasets covering news media and long- and short-form online discussions. The resulting model, called Mformer, outperforms existing approaches on the same domains by 4–12% in AUC and further generalizes well to four commonly used moral text datasets, improving by up to 17% in AUC. We present case studies using Mformer to analyze everyday moral dilemmas on Reddit and controversies on Twitter, showing that moral foundations can meaningfully describe people's stance on social issues and such variations are topic-dependent. Pretrained model and datasets are released publicly. We posit that Mformer will help the research community quantify moral dimensions for a range of tasks and data domains, and eventually contribute to the understanding of moral situations faced by humans and machines.

1 Introduction

Recent years have witnessed a growing interest in the study of moral content on social media. Many online discussions have a tendency to reflect aspects of morality, and researchers thus far have aimed to study how and to what extent moral dimensions vary throughout this vast domain. One particularly influential framework in the analysis of such content is moral foundations theory (MFT), which maps morality to five fundamental psychological dimensions called "moral foundations": *authority*, *care*, *fairness*, *loyalty* and *sanctity* (Haidt and Joseph 2004; Haidt 2012). Prior research suggests that the variation in moral sentiment within and across cultures can be attributed to differences in the way these cultures realize and value each moral foundation. Notable works, including those on vaccine hesitancy (Weinzierl and Harabagiu 2022), social norms (Forbes et al. 2020) and news story framing (Mokhberian et al. 2020), have used this dimension-mapping approach to uncover large-scale patterns of moral belief and judgment.

As human labeling does not scale to the size of modern corpora, MFT-based studies of online moral content must rely on tools to automatically detect moral foundations in text. However, existing methods, especially word count programs based on human-crafted lexicons, are surprisingly lacking in their consistency and ability to generalize to different domains (see Figure 1 for an illustration). Variations across these methods can lead to significant changes in downstream findings based on such measurements. In this work, we propose Mformer, a Moral foundations classifier based on transformers fine-tuned on datasets from diverse domains, which is released publicly.[1] Compared to a set of current approaches, we find that simply using diverse datasets to fine tune works surprisingly well—we observe that Mformer consistently achieves better accuracy on several datasets, with a relative AUC improvement of 4–17%. Through two case studies involving moral stories on Reddit and controversies on Twitter, we demonstrate the effectiveness of Mformer in explaining non-trivial variations in people's moral stances and judgments across many social issues. The main contributions of this work are as follows:

- We introduce Mformer, a moral foundations classifier based on a fine-tuned large language model on text data from diverse domains (Section 4).
- Through an in-depth analysis of word count programs, we show why and how they tend to fall short in labeling moral foundations in text (Section 3). On the other hand, Mformer consistently performs the best across several in- and out-of-domain datasets (Sections 4 and 5).
- We demonstrate the utility of moral foundations in text analysis through two case studies involving (i) moral stories and judgments and (ii) stance toward several controversial topics (Section 6). We highlight the difference between downstream conclusions resulting from word count and those from Mformer. This suggests that many prior findings relying on MFT measurements may warrant further scrutiny.

[1] https://github.com/joshnguyen99/moral_axes

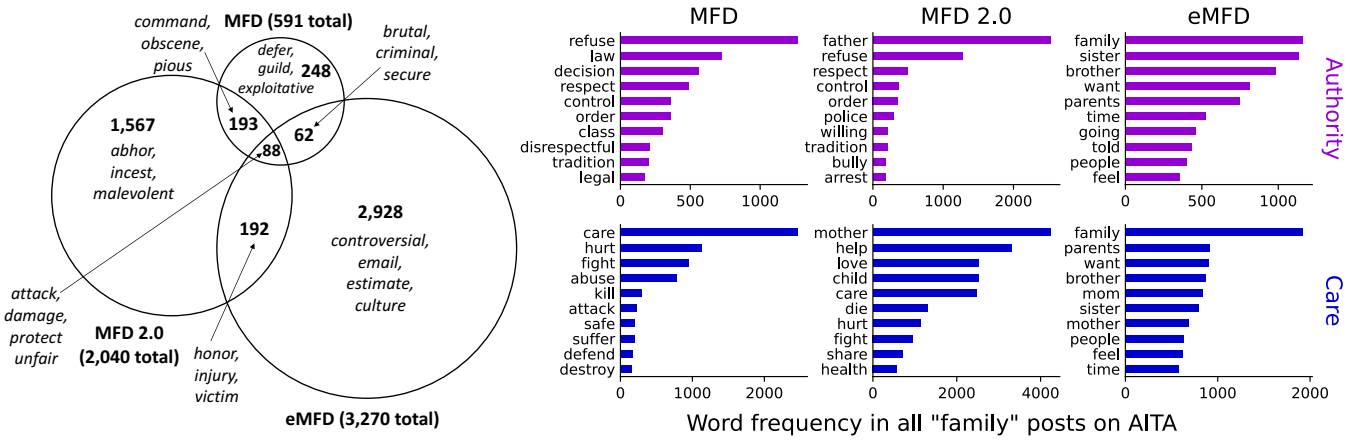


Figure 1: Three existing lexicons—MFD, MFD 2.0, and eMFD—used for word count in detecting moral foundations. *Left*: Venn diagram depicting the sizes of these lexicons with example words. *Right*: 10 most popular words for two moral foundations (*authority* and *care*) in each lexicon that are found in 6,800 `r/AmItheAsshole` posts of the topic *family*. See Section 3.

2 Related Work

Our work builds upon the literature on defining key moral dimensions, detecting them in text, and measuring a variety of patterns pertaining to morality within large corpora.

2.1 Moral Foundations Theory

Moral foundations theory, developed in psychology, proposes a taxonomy over what are called "moral intuitions" (Haidt and Joseph 2004; Haidt 2012). In an attempt to explain the similarities and differences in moral judgment across cultures, Haidt and colleagues postulated key categories of intuitions behind people's moral judgments. These so-called "moral foundations" fall into five spheres: *authority*, relating to traits such as deference to higher authorities to maintain group cohesion; *care*, upholding the virtues of nurturing and protection; *fairness*, involving equal treatment and reward; *loyalty*, relating to prioritizing one's group and alliances; and *sanctity*, including intuitions about maintaining the sacredness of the body and avoiding moral contamination. Anthropological evidence suggests these foundations are universal, although which traits constitute a virtue or a vice vary across cultures (Graham et al. 2013).

Most studies using MFT have focused on characterizing political ideology, especially in the U.S. context. For example, pertaining to the foundations *care* and *fairness*, liberals often support the commitment to justice and actions that uphold equality and minimize suffering. Conservatives, on the other hand, tend to value all five foundations somewhat equally, additionally upholding virtues such as loyalty to one's country and respecting authority and order (Graham, Haidt, and Nosek 2009; McAdams et al. 2008; Haidt and Graham 2007; van Leeuwen and Park 2009). The theory has also been used to characterize the differences in moral sentiment surrounding socially significant topics such as stem cell research (Clifford and Jerit 2013), vaccine hesitancy (Amin et al. 2017), euthanasia, abortion, animal cloning, and same-sex marriage (Koleva et al. 2012).

2.2 Automatically Detecting Moral Foundations

A requirement for large-scale studies of MFT is the automatic detection of moral foundations. While the gold standard for moral foundation labeling is human annotation, this requires extensive training and evaluation, which does not feasibly scale to large text collections. We categorize the automatic scoring of moral foundations into word count, embedding similarity, and supervised classification methods.

Word count methods rely on a human-crafted lexicon, which is a mapping from words to foundations. When scoring a document, each token contained in the lexicon counts as one occurrence of the foundation that it maps to. Three lexicons, often called moral foundations dictionaries (MFDs), have been extensively used in the literature: MFD (Graham and Haidt 2012), MFD 2.0 (Frimer et al. 2019), and extended MFD (Hopp et al. 2021); we further review these in Section 3. Word count methods resemble LIWC, a popular program that scores psychological and emotional features in text (Tausczik and Pennebaker 2010).

Embedding similarity methods define the relevance between a document and a moral foundation as the cosine similarity between their word embeddings such as word2vec (Mikolov et al. 2013). In particular, each foundation is represented by the average embedding of its "seed words," and similarly each document is the average of its tokens. Mokhberian et al. (2020), for example, used this method to score moral foundations as features of news articles. Similar to Rocchio classification (Manning, Raghavan, and Schütze 2008), embedding similarity essentially relies on the centroid of a set of seed words and is well-known to have limited expressivity. Furthermore, choosing a good set of seed words remains a major practical challenge.

Supervised classification methods treat the detection of moral foundations as a text classification task. For example, Hoover et al. (2020) annotated a dataset of tweets and used it to train support vector machines for classification. Similarly, Trager et al. (2022) fine-tuned BERT (Devlin et al. 2019), a large language model (LLM), using annotated social media

data for the same task and reported state-of-the-art performance. However, these models may not generalize well to new datasets (Liscio et al. 2022). Recently, Guo, Mokhberian, and Lerman (2023) trained moral foundation classifiers with a new loss function to incorporate domain variations. In comparison, our approach uses a standard LLM architecture trained on multiple data domains. This approach is arguably more straightforward as it requires no additional hyperparameters, does not demand explicit one-hot encoding of the domains, and thus offers improved adaptability when a new domain enters into the training dataset.

Finally, some prior work made a distinction in the polarity of moral foundations, i.e., whether an example portrays a virtue or a vice of a foundation. As we explain later in Section 4.1, there are several conceptual and practical concerns for this approach. Thus, we decide to score only moral foundations, irrespective of polarity, in this work.

2.3 Analyses of Text

A growing body of work has adopted MFT to the analysis of moral rhetoric from online media. For example, Mokhberian et al. (2020) studied the relationship between story framing and political leaning of several news sources. The study found that on topics such as immigration and elections, conservative-leaning sources tend to focus on the virtues of *sanctity* such as austerity and sacredness, while liberal-leaning sources emphasize the condemnation of the vices of *sanctity* like dirtiness and unholiness. Hopp, Fisher, and Weber (2020) analyzed movie scripts and identified relevant foundations in movie scenes exhibiting moral conflicts. Using network-theoretic methods alongside moral foundations, the authors were able to construct communities of characters with specific shared moral characteristics. In argument mining, Kobbe et al. (2020) proposed to use moral foundations in the automatic assessment of arguments. The study found a significant correlation between moral sentiment (the presence of moral foundations) and audience approval of an argument. In another line of work, Forbes et al. (2020) and Ziems et al. (2022) used MFT to categorize collections of moral "rules of thumb" collected from large crowds and assessed LLMs' moral sentiment predictions.

Empirical results based on moral foundations scoring currently face two important limitations. First, there is no unified and highly accurate method on which researchers rely to label their text corpora. Second, and as a result, conclusions drawn from prior studies likely suffer from low-quality scoring and may vary based on what method is chosen. A solution sufficiently addressing these two challenges will allow MFT-based analyses to scale and facilitate their reproducibility, comparison, and generalization.

3 Limitations of Moral Foundations Dictionaries

Word count methods based on lexicons remain the most popular for automatically detecting moral foundations, often serving as the default choice among researchers. Here we detail how they work and several of their limitations when applied to text corpora.

Scoring via word count Each lexicon, called a moral foundations dictionary (MFD), contains a list of words and the foundations they represent. For example, the word "deceive" can be mapped to the foundation *loyalty*. A document to be scored is first tokenized; then, for each token that is also in the MFD, the count for the foundation that it is mapped to is incremented. For instance, if the token is "deceive," the count for *loyalty* is increased. These methods are easy to implement, involve no model training, and are interpretable since they directly show what words signify a foundation.

Available lexicons Three versions of the MFD have been used extensively in the literature. The first MFD, released as a dictionary to be used in LIWC-like programs, contains nearly 600 words (Graham and Haidt 2012). Later, a new version with over 2K words called MFD 2.0 (Frimer et al. 2019) was released in which the creators extended their expert-crafted word lists by querying a word embedding (Mikolov et al. 2013) for similar terms. Another recent variant, called the extended MFD (eMFD, 3.2K words) (Hopp et al. 2021), contains one-to-many mappings associated with moral foundation weights, between 0 and 1, for each word. We list some example words of these MFDs in Table A.1 in the appendix and show some overlap between them in Figure 1.[2]

Limitations We want to understand the different MFD versions and what they each capture, but have since identified other limitations that will affect downstream measurements and interpretations.

First, these lexicons have a fixed and limited vocabulary. See Figure 1 (left) above for some example words and a visualization. Most of the words in these lexicons are supposedly morally relevant, but their overlap is surprisingly small. For instance, the 281 words common to both MFD and MFD 2.0 only account for 47.5% of MFD and 13.8% of MFD 2.0. In eMFD, 89.5% of the words do not appear in either MFD or MFD 2.0 at all. Such lack of consensus in expert- and machine-created lexicons is concerning, and one can rightfully question whether the resulting scores are reliable.

Second, when using the MFDs, it is unclear how word variations should be handled. For example, the MFD 2.0 explicitly contains the verb "desecrate" and its inflected forms "desecrates," "desecrated" and "desecrating." However, this is not the case for the verb "deify," which is in the lexicon while "deified" is not. A direct but possibly non-exhaustive way to address this is to lemmatize all tokens in a document before performing a lexicon lookup. Another related issue is how to disambiguate the parts of speech of some words in these dictionaries. For example, given the word "bully" in the MFD 2.0, should it be counted when it is a verb, noun, or even an adjective (which, in this case, means "excellent")? Researchers have attempted at disambiguation (Rezapour, Shah, and Diesner 2019), but accounting for word senses alone does not mitigate the drawbacks of lacking inflections or the incompleteness of the vocabulary.

Third, longer documents tend to have a higher chance of having a dictionary match. In Appendix Figure C.8 (top panel), we show that the number of words found in each

[2]The appendix is found at https://arxiv.org/abs/2311.10219.

dictionary is highly correlated with how many words each text input has (at r=0.59, 0.72, and 0.98 respectively for MFD, MFD 2.0, and eMFD). When foundation scores are normalized by the input length, such strong positive relationships disappear (Figure C.8, bottom panel), suggesting that length-normalized scoring might be preferable if one has to use dictionary-based methods. On the other hand, there are examples for which the detected moral foundation is due to one matching word in a long post—Appendix C.1 shows an example that triggers the foundation *authority* because the word "refuse" appears once in a 140-word long post.

Finally, the hand coding of words to moral foundations by experts is a source of personal subjectivity and social bias. In the same analysis using Reddit posts, we discover some problematic associations. Figure 1 (right) presents the most frequently matched words that are related to two foundations: *authority* and *care*. Using MFD 2.0, some associations such as "father" with *authority* and "mother" with *care* appear to reflect the bias of the lexicon creators when assigning moral intuitions to very general familial roles. Appendix C.2 contains a systematic comparison of foundation scores for posts containing "father" and those with "mother," showing that posts with "mother" scores higher for *care* (MFD and MFD2.0) and *loyalty* (MFD), and that posts with "father" score higher for *authority* (MFD 2.0) and lower on *care* (MFD 2.0). The same figure also shows that Mformer does not suffer from the same bias.

These limitations are not restricted to moral foundations. Within sentiment analysis, where lexicons such as LIWC (Tausczik and Pennebaker 2010) are used, the same pitfalls are especially illuminating. First, word count explicitly ignores the context-dependent nature of language by relying on (normalized) frequency as a direct proxy to sentiment (Pang and Lee 2008; Puschmann and Powell 2018), and by treating polysemous words equivalently (Schwartz and Ungar 2015). Second, since lexicons are static, the validity and accuracy of this method highly depend on its input's domain (González-Bailón and Paltoglou 2015). And third, word count can be shown to predict sentiment more poorly than simply word presence, suggesting that the length of an input may influence the overall score more substantially (Pang and Lee 2008). Machine learning approaches, especially with the advent of LLMs, can overcome these challenges through their ability to learn contextualized patterns and superior cross-domain generalizability.

Overall, dictionary-based moral foundation scoring seems fragile, lacks consensus in what they capture, and has inherent biases in social stereotypes. We caution the use of manually curated lexicons without further scrutiny, and recommend examining their coverage and accuracy before interpreting aggregate results. Recognizing that developing a good lexicon takes significant effort and is difficult to get right, we adopt a data-driven approach for moral foundation scoring which avoids these limitations and can account for the nuances in language that exist beyond individual words.

4 Constructing and Evaluating Mformer

In this section, we describe Mformer, a language model fine-tuned from a wide range of data to score moral foundations in text. We introduce the datasets Mformer is trained on in Section 4.1. Then we describe the training procedure along with some baselines (Section 4.2). Finally, we present evaluation details that highlight Mformer's efficacy (Section 4.3).

Source	Twitter	News	Reddit	Total
Data Period	'10–'17	'12–'17	'20–'21	–
# Examples	34,987	34,262	17,886	87,135
Avg. # Tokens	19.3	28.0	41.7	27.3
# Annotators	854	13	27	–
% Authority	33.4	24.9	19.2	27.1
% Care	40.6	24.8	26.5	31.5
% Fairness	35.9	24.2	29.5	30.0
% Loyalty	31.1	24.4	11.1	24.4
% Sanctity	22.3	19.9	9.8	18.8

Table 1: Three moral foundations datasets used to develop Mformer.

4.1 Datasets

We first describe the dataset used to train and evaluate moral foundation classifiers. We combine three publicly released, high-quality data sources labeled with moral foundations.

Twitter (Hoover et al. 2020) This dataset contains 34,987 tweets encompassing seven "socially relevant discourse topics": All Lives Matter, Black Lives Matter, 2016 U.S. Presidential election, hate speech, Hurricane Sandy, and #MeToo. Annotators were trained to label the tweets with moral foundations and their sentiments (virtue and vice), with at least three annotations per tweet. We keep all tweets in this dataset for our use and determine that a tweet contains a foundation f if at least one annotator labeled it with f. Further, for each foundation, we merge the labels for its virtue and vice into one: e.g., the raw labels "purity" and "degradation" are mapped into the same foundation *sanctity*.

News (Hopp et al. 2021) This dataset was used to construct the eMFD lexicon (cf. Section 2.2). The authors pulled 1,010 news articles, most of which on politics, from the GDELT dataset (Leetaru and Schrodt 2013) and employed workers to label these articles with moral foundations. Specifically, each annotator was assigned a foundation-article pair and then asked to highlight all sections in the article that contain this foundation. We segment every article into sentences and assign a moral foundation f to a sentence if any part of it is contained within a highlighted section labeled with f. This yields 32,262 instances in total.

Reddit (Trager et al. 2022) This dataset contains 17,886 comments on 12 different subreddits roughly organized into three topics: U.S. politics, French politics, and everyday moral life. In annotation, the authors separated the foundation *fairness* into two classes: *equality* (concerns about equal outcomes for all individuals and groups) and *proportionality* (concerns about getting rewarded in proportion to one's merit). Another label, *thin morality*, was defined for cases in which moral concern is involved but no clear moral foundation is in place. We merge both *equality* and *proportionality*

into their common class *fairness* and consider *thin morality* as the binary class 0 for all foundations, which results in the same five moral foundation labels. Finally, for each comment, we assign a binary label 1 for foundation f if at least one annotator labeled this comment with f.

A profile of the datasets We combine the three sources—Twitter, News, and Reddit—into one dataset, yielding 87,135 instances with 2.4M tokens. Table 1 presents summary statistics. Each example has on average 27.3 tokens, with Reddit comments the longest (41.7 tokens) and tweets the shortest (19.3 tokens). The foundations *care* and *fairness* have the most positive instances in total, each with at least 30% of the dataset. Among the three sources, tweets tended to contain more foundations than Reddit comments. For example, over 31% of tweets contain *loyalty* while only 11.1% of Reddit comments do. Finally, for each foundation, we split this dataset into a training and test set with ratio 9:1, stratified by that foundation. In Appendix D, we describe the datasets in more detail, including their annotation scheme and agreement rate, and how label disagreement and train-test splitting are handled. Finally, there exist other datasets labeled with foundations, such as *covid* and *congress* used by Guo, Mokhberian, and Lerman (2023). We choose not to include them due to their significantly smaller size—in the 1–2,000 range rather than 15,000+.

Capturing moral foundation polarity Some prior work has additionally considered *polarity*, i.e., whether a text instance conveys a *virtue* or a *vice* of a moral foundation, resulting in ten classes (two for each foundation). In this work, we decide against this approach—instead only aiming to score the relevance of a foundation regardless of polarity—for three reasons. First and conceptually, virtues and vices are very loosely-defined terms whose perception is subject to cultural differences (Graham et al. 2013, see §2.4.4 for an example of *authority*). Second, while some previous work has treated the virtue/vice distinction as a sentiment analysis task (Hopp et al. 2021), we believe this is somewhat naïve since it lacks a theoretical justification. Third and operationally, the assignment of virtues or vices by human annotators is another source of noise on top of the noise in moral foundation labels. This is coupled with the fact that not all available datasets/lexicons capture this polar distinction. We do not argue that virtues and vices are irrelevant; rather, we believe they deserve a more thorough theoretical and practical treatment, which is beyond this work's scope.

4.2 Moral Foundations Classifiers

Mformer LLMs have achieved state-of-the-art performance across a range of NLP benchmarks. Our work is not the first to use LLMs for this task; for example, Trager et al. (2022) fine-tuned BERT (Devlin et al. 2019) to create moral foundation classifiers. However, we note that prior work primarily focused on setting up a baseline for future work. As such, a careful treatment of the fine-tuning process and evaluation is necessary to substantiate the adoption of such methods.

We choose the RoBERTa-base architecture (Liu et al. 2019) with 12 self-attention layers for this task. Each document is tokenized and then two special tokens, `<s>` and `</s>`, are added to the beginning and end of the document, respectively. A classifier module follows the final attention layer, where the 768-dimensional embedding of the `<s>` token goes through a fully connected layer with 768 neurons followed by $\tanh$ activation. Finally, this is linearly mapped to a two-dimensional output vector and then converted to probabilities via a softmax layer. In fine-tuning RoBERTa, we find the optimal learning rate and the number of training epochs by performing a grid search. We end up with five binary classifiers, each of which outputs a score between 0 and 1 for every input text. We call the final fine-tuned models Mformer, for Moral foundations using transformers. More training details are found in Appendix E.2.

Baselines For comparison, we consider as baselines all methods described in Section 2.2: *word count*, *embedding similarity*, and *supervised classifiers*.

For word count, we score a document based on the description in Section 3. We experiment with three lexicons: MFD, MFD 2.0, and eMFD. For MFD and MFD 2.0, we increment the foundation count by one every time its example word is encountered and then divide the count by the total number of tokens. This represents the frequency with which the foundation is found among the tokens. For eMFD, since the lexicon contains soft counts between 0 and 1, every time a word in the dictionary is found we add all scores to their corresponding foundations. Then, the five-dimensional vector of foundation scores for the document is normalized by the number of tokens that match the eMFD's entries. For all three lexicons, the foundation scores are in $[0, 1]$. More detail is found in Appendix A.

For embedding similarity, with each foundation f and a document d, the score for d is defined as the cosine similarity between the embedding vectors for f and d. To encode f and d, we use the GloVe embedding (Pennington, Socher, and Manning 2014), specifically the "Twitter" 200-dimensional version. The vector representation for f is defined as the average of the vectors for the "seed words" that represent f. Similarly, the vector for d is the average of the vector representations of all of its tokens. The range for foundation scores is $[-1, 1]$. For more detail, including the seed words describing each foundation, see Appendix B.

Finally, for supervised classifiers, we train a logistic regression model on a range of sparse and dense embeddings. We find that, unsurprisingly, the embedding with the best performance is Sentence-RoBERTa, which is based on RoBERTa fine-tuned for sentence similarity (Reimers and Gurevych 2019). In Appendix E.1, we provide more details of logistic regression and compare it with support vector machine as used in previous work (Hoover et al. 2020).

Alternative to binary classifiers Mformer is a collection of five binary classifiers each for one moral foundation and the corresponding set of RoBERTa weights. We also consider a weight-shared variant in which only one model is used but the final classification layer contains five neurons, each followed by a sigmoid activation. In other words, this multi-label model outputs five binary probabilities simultaneously predicting each foundation. Compared to Mformer, multi-label RoBERTa requires less storage and training resources. However, we find that this model performs uniformly worse than Mformer, achieving 10.7–19.3% lower test AUC than

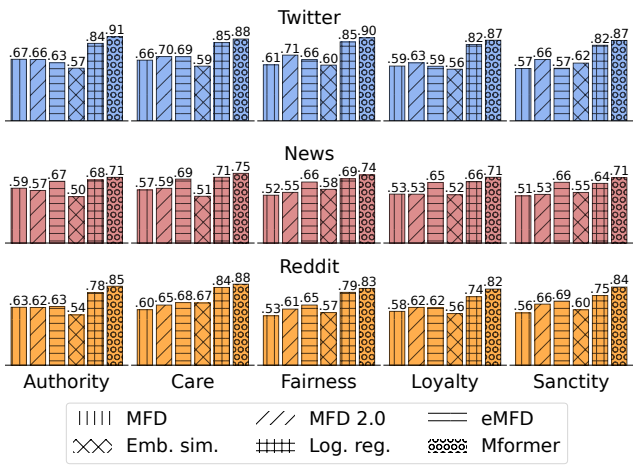


Figure 2: AUC on the Twitter (top row), news (middle row) Reddit (bottom row) portions of the test set for five moral foundation scoring methods: MFD, MFD 2.0, eMFD, embedding similarity, logistic regression, and Mformer.

its binary classification counterparts (see Appendix E.3).

4.3 Evaluation

We evaluate classification methods presented in Section 4.2 using the hold-out test set in Section 4.1. The results show that Mformer outperforms all existing methods in scoring all foundations, often by a considerable margin.

Evaluation metric It is worth noting that this dataset is multi-labeled: each instance contains between zero and five foundations. Our goal is to build five classifiers each predicting the binary label of each foundation given an input. Two considerations are taken into account. First, all classifiers described in this section output a "score" representing the likelihood that a foundation exists in an input. Second, as shown in Table 1, the dataset is unbalanced for all foundations with the percentage of positives being as low as 18.8%. A suitable metric should be *threshold-free* (it considers the scores and not just binary predictions), *scale-invariant* (it considers prediction scores ranked on any scale), and take into account *unbalanced class prior*. We therefore choose the area under the ROC curve (AUC) for evaluation. A useful statistical property of this metric is that the AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. While the range of the AUC is $[0, 1]$, a realistic lower bound is 0.5 which represents a classifier that randomly guesses positive half of the time (Fawcett 2006).

In-domain evaluation We score all test examples using the methods described in Section 4.2 and report the test AUC in Table E.1 in the appendix. We find that embedding similarity shows the worst performance with AUC between 0.51 and 0.59 for all foundations, only slightly better than random guessing. Simple word count methods surprisingly perform better than embedding similarity, with each updated version of the MFD tending to improve from the previous. Logistic regression models further improve from word count, achieving the AUC between 0.76 and 0.81 for the Sentence-BERT embedding. The highest recorded test AUC for all foundations is by Mformer, where all foundations achieve an AUC between 0.83 and 0.85, a relative improvement of 4–12% from logistic regression. When comparing the performance across foundations, we find that *care* and *fairness* are the easiest to score (both AUC = 0.85), while *loyalty* and *sanctity* are the more difficult but not sizably (both AUC = 0.83).

Since the test sets are merged from three sources—Twitter, News, and Reddit—we examine the performance of these classifiers on each subset in Figure 2. The findings are similar: embedding similarity performs the worst, followed by word count (which gets better each newer version of the lexicon), then by logistic regression, and Mformer remains the best. Of all domains, tweets are the easiest to score: the AUC for *authority* and *fairness* goes up to 0.91 and 0.9, respectively. In contrast, sentences taken from news articles are the hardest: the AUC for all foundations only ranges from 0.71 to 0.75, with the lowest for *loyalty*. We suspect that the relatively low performance on news sentences is because of the way they were labeled: In (Hopp et al. 2021), annotators highlighted the *sections* of an article containing a foundation, while we subsequently process these sections using *sentence segmentation*. On the other hand, every tweet and Reddit comment was independently labeled for every foundation, which explains the quality of their labels.

Given its superior performance, we adopt Mformer as our final classifier. In Figure 3, we show the highest-scoring examples for every foundation. The scores displayed in the right-hand column suggest that a lot of examples contain more than one moral foundation. For example, according to Mformer, the Reddit comment "Wage slavery is indeed disgusting. Stay strong and safe comrade I wish you the best of luck. Educate agitate organize" conveys *care*, *loyalty*, and *sanctity* with very high confidence. The coexistence and interplay of these foundations suggest complex moral constructs, which we will examine later in this paper.

5 Mformer: Out-of-Domain Evaluation

As shown in Section 4, Mformer demonstrates superior predictive performance to existing methods on the test set. Here, we further highlight that Mformer also generalizes well to other data domains—one in the psychology literature and three in NLP—without any further fine-tuning. Specifically, in this section, we describe each dataset in detail and discuss Mformer's performance presented in Figure 4. Cross-domain evaluation of moral foundations classifiers has been studied in Liscio et al. (2022); however, the "domains" in their work are only restricted to the seven topics in the Twitter dataset (described in Section 4.1). Given the positive results recorded, we emphasize the potential of Mformer to be adopted to many analyses of moral rhetoric based on MFT without the costly training of a new model.

Moral foundation vignettes (VIG) (Clifford et al. 2015) This dataset contains 115 vignettes designed by the authors to assess humans' classification of moral foundations. Each vignette is a short description of a behavior that violates a foundation. An example for *fairness* is "You see a politician using federal tax dollars to build an extension on his home."

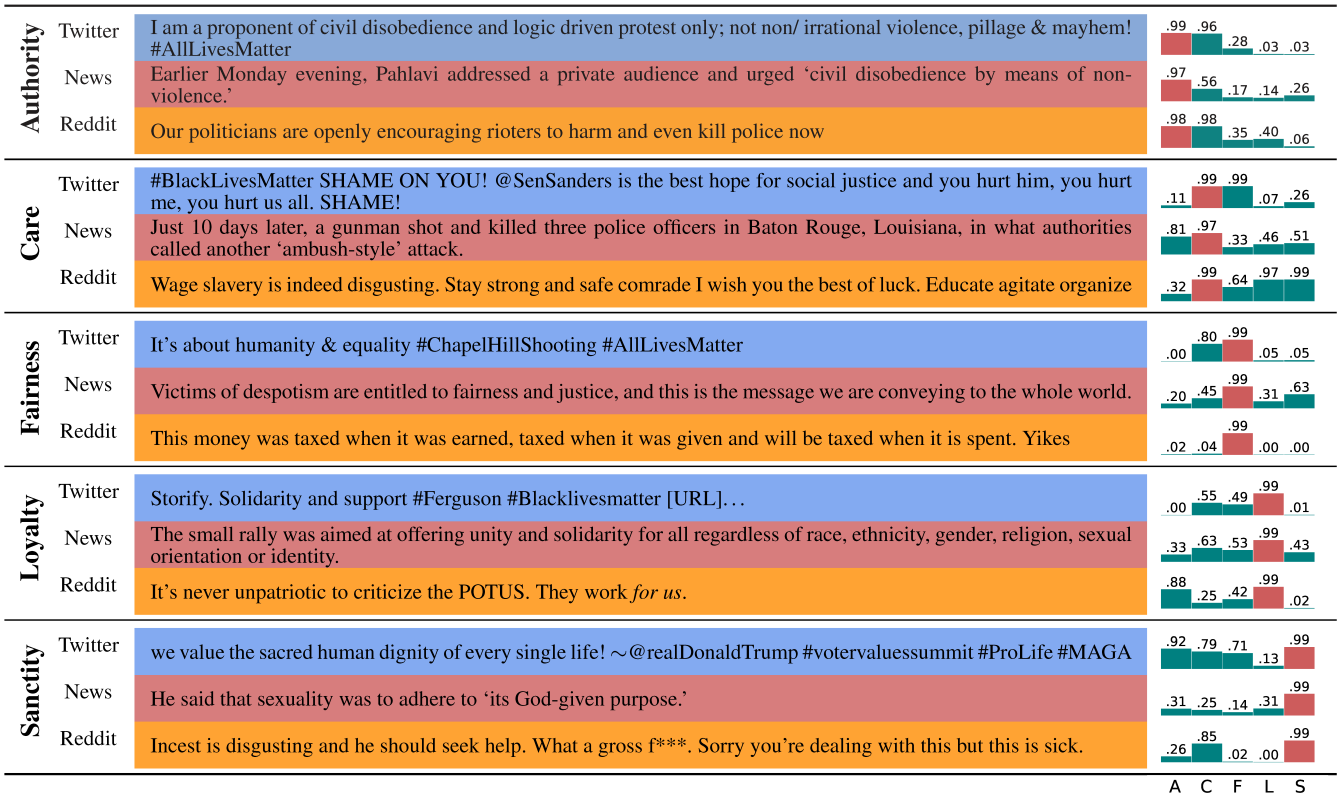


Figure 3: Highest-scoring test examples for each foundation. The right-hand bar charts show Mformer's predicted scores.

As presented in Figure 4, Mformer performs very well on this dataset, achieving an AUC of 0.95 for *authority* and higher than the second-best method, logistic regression, by 7–15%. The only surprising exception is *loyalty*, on which Mformer achieves an AUC of 0.75, slightly lower than logistic regression of 0.76 and equal to embedding similarity. Upon inspection, we find that some examples of *loyalty* tend to be misclassified as *authority* like the following vignette: "You see a head cheerleader booing her high school's team during a homecoming game."

Moral arguments (ARG) This dataset contains 320 arguments taken from two online debate platforms (Wachsmuth et al. 2017). Kobbe et al. (2020) subsequently labeled each argument with moral foundations. On this dataset, we also observe very good results for Mformer with all AUC between 0.81 and 0.86, the highest among all methods and up to 17% higher than the AUC for logistic regression. We find *authority* and *sanctity* relatively more difficult to classify. Some instances of *authority* tend to be confused with *care*; e.g., "Some kids don't learn by spanking them.So why waste your time on that, when you can always take something valuable away from them." This is also observed for arguments containing *sanctity*—see Appendix G for an example.

Social chemistry (SC) (Forbes et al. 2020) This dataset contains 292K moral rules-of-thumbs (RoTs) labeled with moral foundations, social judgment and others. We use the test set and score all of its 29K instances. For Mformer, the AUC ranges between 0.70 and 0.80—highest among all methods—with specifically high AUC for *loyalty*. We also find that logistic regression comes close to Mformer, and is much better than word count methods which often perform marginally better than chance. As we explain in more detail in Appendix G, the relatively low performance of Mformer, compared to that observed in VIG or ARG, may be attributed to this dataset's high level of label noise. As an example, the following RoT is predicted with a very high score for *care* but does not contain this ground-truth label: "People should temper honesty with compassion, especially when it comes to family."

Moral Integrity Corpus (MIC) This dataset contains 99K annotated RoTs derived from 38K responses to questions on Reddit (Ziems et al. 2022) by chatbots to facilitate the study of their moral biases. We use the test set with 11K examples for evaluation. Similar to SC, the AUC for Mformer on this dataset, ranging from 0.65 to 0.75, is lower than that on VIG or ARG, but remains the highest among all methods. We also suspect that this is largely due to label noise in the dataset as the RoTs were labeled similarly to those in SC. For instance, this RoT is predicted with a high score for *fairness* but does not contain this label: "It's wrong to fight in an unjust war."

6 Studying Moral Dilemmas and Controversies using Mformer

So far, we have established that current methods used to score moral foundation may actually perform not much bet-

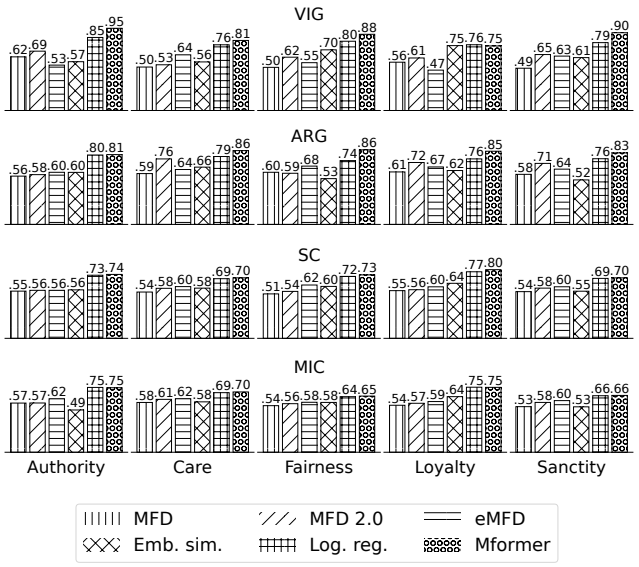


Figure 4: AUC on four external datasets for six moral foundations scoring methods: MFD, MFD 2.0, eMFD, embedding similarity, logistic regression, and Mformer.

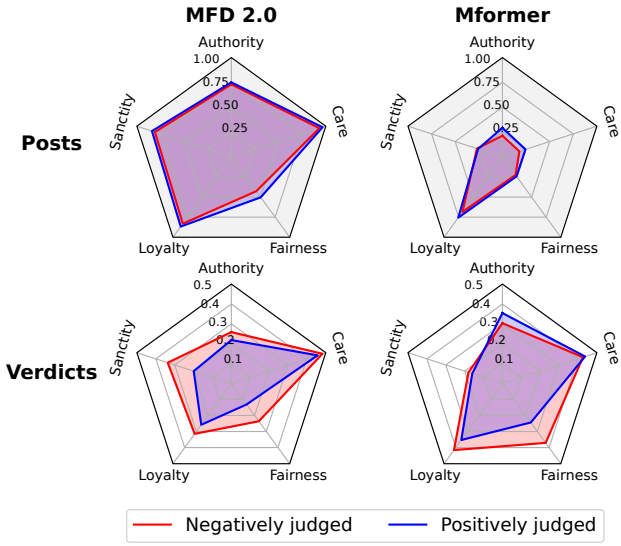


Figure 5: Posts (top) and verdicts (bottom) in the (*family*, *marriage*) topic pair on AITA. Each number in a radar plot indicates the proportion of posts (or verdicts) that contain the corresponding moral foundation. The moral foundations are detected by two methods: MFD 2.0 and Mformer. Red (resp. blue) indicates negative (resp. positive) verdicts.

ter than chance, and we have advocated for the adoption of Mformer based on its good performance across a number of domains without any further fine-tuning. In this section, we present some case studies, partly replicating some previous work, that highlight Mformer's efficacy in discovering patterns of morality in several socio-political domains.

6.1 Moral Dimensions in Everyday Conflicts

Recently Nguyen et al. (2022) analyzed over 100,000 moral discussions on `r/AmItheAsshole` (AITA), where users post an interpersonal conflict they have experienced and ask the community to judge if they are in the wrong. The authors used topic modeling to find the most salient topics of discussion in this community and found that topics and topic pairs are a robust thematic unit over AITA content. They then used the MFD 2.0 to label all posts and verdict comments to examine the patterns of framing and judgment pertaining to moral foundations across all topics and topic pairs.

Here we replicate the same study, this time using Mformer as a moral foundation labeling method instead of MFD 2.0. Our aim is to examine whether the findings in the previous study still hold when a better classifier is used. For more detail on the setting, see Appendix H.1. Similar to Nguyen et al. (2022), we calculate the *foundation prevalence*—defined as the proportion of posts/verdicts that contain each moral foundation—in each topic to examine the relative importance of these five moral foundations within every sphere of moral discussion.

Figure 5 presents the radar plots for foundation prevalence among all posts and verdicts in the (*family*, *marriage*) topic pair. Using MFD 2.0, we find that all foundations except *fairness* are salient among posts in this topic; however, the results by Mformer indicate that only *loyalty* is dominant. For verdicts, while the results by Mformer agree with the previous finding that the foundation *care* is significant, we also find that it is *loyalty*, not *sanctity*, that is of major concern among these judgments. Differences in prevalence also hold for most of the 47 topics found in that study, which we present in Figures H.16 and H.17 in the appendix.

These results highlight that important findings are subject to change when researchers use different scoring methods. We believe that Mformer is a good candidate for adoption as it performs better than all existing methods across numerous benchmarks (Sections 4 and 5) and is robust to its internal binary thresholds (see Figure H.15 in the appendix).

6.2 Moral Dimensions in Opposing Judgments

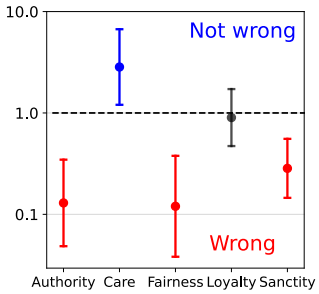
Here we present another study using AITA content. In contrast to Nguyen et al. (2022), who only looked at posts and their verdicts (i.e., highest-scoring judgments), we aim to analyze *all judgments* within *one post* to find any systematic differences in conflicting judgments—those that claim the author is in the wrong and those who think otherwise. To do so, we filter the dataset to contain only "controversial" posts with at least 50 judgments, which are split somewhat equally between the positive and negative valence. This yields 2,135 posts accompanied by 466,485 judgments. A detailed setting can be found in Appendix H.2.

We use Mformer to score every post and judgment on five moral foundations. To convert the scores to binary labels, we set the highest-scoring 20% of the posts to contain that foundation; the same applies to judgments. In other words, a post (or judgment) is said to contain foundation *loyalty* if it scores higher than at least 80% of all posts (or all comments) on *loyalty*. This rather high threshold is motivated by our

AITA for inviting my atheist friends only for the wedding party and not the ceremony?

I am getting married in a few months. We have a few atheist friends who have a general dislike for religion so I thought it would be for the best to only invite them to the party. My future wife and both of our families are Roman Catholic. Catholic ceremonies are *long*. […] If you are not a believer, it is boring **. […] I thought I'll spare them the trouble.

Anyway, one of my friends brought it up, saying I forgot to put the time and the church where the ceremony will take place on the invitation. So I said that she doesn't have to come to the religious part and just come for the wild party part. She got confused, I explained and she got annoyed to say the least. Said she isn't heartless and would sit through the 2/3 to see me get married in the 1/3. I tried to explain again & fix things but I felt guilty the whole time.



fairness

YTA (mildly). You were trying to be considerate but in the process basically told your friends you thought they weren't considerate. Ceremonies are boring for everyone except for the bit that matters (the actual getting married part). Why not make it optional for everyone? That way no one feels excluded. (not to late to amend)

authority, loyalty

YTA – don't assume what people want. Give them the option to go. No one in their right mind is going to disgrace your religion at your wedding

care, loyalty, sanctity

NTA, you did what you thought was best & tried to put their best interests at heart taking into consideration their religious stance. Even though you ended up being wrong you were still trying to be considerate. Besides, everything worked out; have a very good wedding! :)

Figure 6: Example controversial thread on AITA. *Top left*: the post, including its title in bold and body text. *Top right*: odds ratios and 95% CIs between the presence of a moral foundation in a judgment and the judgment's valence. Values above the dashed horizontal line indicate that a foundation is associated with positive valence (i.e., "NTA" or "NAH"), while values below the dashed line indicate an association with a negative (i.e., "YTA" or "ESH") judgment. *Bottom*: three judgments for this post. The foundations contained in each judgment are annotated at the top.

striving for high precision at the expense of recall. For each post and a foundation f, we calculate the odds that a positive ("not wrong") judgment contains f, compared to the odds that such a positive judgment does not contain f.

Figure 6 displays an example controversial thread on AITA. The post received 397 positive (the author is "not wrong") and 471 negative ("wrong") judgments. The odds ratio plot at the top right suggests some distinct patterns: those that think the author is not wrong are 1.6 times more likely to focus on *care* (OR=1.57, 95% CI=[1.08, 2.28]) by arguing that the author is simply looking out for their atheist friends and helping them avoid a religious event they might be uncomfortable with. On the other hand, those that judge the author to be in the wrong are 2.5 times more likely to emphasize the foundation *fairness* (OR=0.40, 95% CI=[0.24, 0.66]) by stating that the author ought to be fair to all guests by inviting them to the ceremony as well. Negative judgments are also more likely to underlie *authority* (OR=0.41, 95% CI=[0.27, 0.63]) and *sanctity* (OR=0.58, 95% CI=[0.43, 0.77]); these judgments often argue that the author should not assume, on behalf of his friends, that they would not want to be at the ceremony just because it is religious and they are not. Finally, we find that the foundation *loyalty* is not significantly associated with positive or negative judgments (OR=0.96, 95% CI=[0.72, 1.27]); it seems that within this situation, the author's loyalty to their friends is not being questioned as much as other moral values.

Not all moral dilemmas give significant findings. Even after filtering less controversial posts, we only find, among the 2,135 controversial threads, 1,136 (53.2%) of them to have significant results where at least one moral foundation is associated with a clear valence of wrong/not wrong. Nevertheless, the results suggest that conflicting judgments for moral dilemmas can be explained by their appeal to different moral foundations, which can be robustly detected by Mformer.

6.3 Moral Dimensions of Different Stances

Stance classification is concerned with determining whether a person is in favor of or against a proposition or a topic (Mohammad, Sobhani, and Kiritchenko 2017). Here we explore the endorsement of moral foundations and its relationship with people's stance toward several controversial topics, as expressed through tweets. This study partially reproduces the analysis in Rezapour, Dinh, and Diesner (2021). We use a dataset of 4,870 tweets across six political topics: *atheism*, *climate change is a real concern*, *Donald Trump*, *feminist movement*, *Hillary Clinton* and *legalization of abortion* (Mohammad et al. 2016). Since we are only interested in polar stances (in favor or against), we remove all instances where the stance was labeled as "none" (i.e., either neutral or irrelevant to the topic). This results in 3,614 tweets in total, with the number of tweets per topic between 361 and 779.

We score all tweets on every moral foundation using Mformer and convert the raw predicted scores to binary labels by setting the highest-scoring 20% of the tweets to contain each foundation. We then replicate the chi-square analysis by Rezapour, Dinh, and Diesner (2021, Section 5.3), the results of which are presented in Table I.3 in the appendix. Apart from some similar findings, such as that no association is found between moral foundations and stance toward *feminist movement*, we find many differences after using Mformer for detecting foundations instead of the MFD in the previous study. For instance, Rezapour, Dinh, and Diesner (2021) found no significant correlation within the topic *Donald Trump*, but our results show that this happens for three foundations: *authority* ($p < 0.001$), *care* ($p < 0.05$)

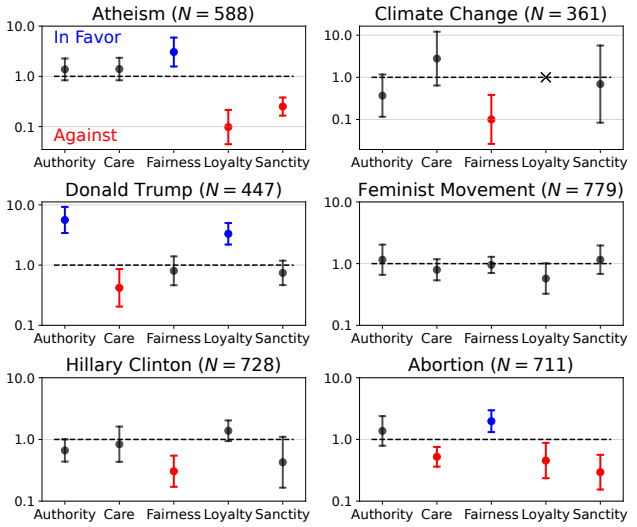


Figure 7: Odds ratios and 95% CIs between the presence of a moral foundation in a tweet and the tweet's stance toward each topic. The "X" mark for *loyalty* in climate change is due to no tweets containing this foundation.

and *loyalty* ($p < 0.001$). We suspect that these differences are largely due to the aforementioned shortcomings of word count methods: Most tweets are short, and so can easily fail to contain lexical entries, leading to zero counts for some foundations and hence false negatives.

Finally, we estimate the effect size of these associations by calculating the odds ratio (OR) between two binary variables: whether a foundation is present in a tweet and whether the tweet's author is in favor of a topic. Figure 7 shows the ORs for each topic over all five foundations. Almost all topics show a clear distinction in moral foundations between tweets in favor of and those against it. For example, Twitter users supporting the *legalization of abortion* are 2 times more likely to appeal to *fairness* (OR=1.98, 95% CI=[1.32, 2.97]), such as in the tweet "Good morning @JustinTrudeau. Do you plan to tell @WadePEILiberal that women on PEI deserve the right to choose? #cdnpoli #peipoli." On the other hand, those against this view are more likely to underlie the foundations *care* (OR=0.52, 95% CI=[0.36, 0.75]), *loyalty* (OR=0.45, 95% CI=[0.24, 0.88]), and, especially, *sanctity* (OR=0.29, 95% CI=[0.15, 0.56]) as in the tweet "U don't have to be a religious to be pro-life. All u have to believe is that every life is sacred." The strong association between the disapproval of abortion and the foundations *sanctity* and *care* is consistent with the finding in Koleva et al. (2012), although this prior work found mixed results when it comes to *fairness* and *loyalty*.

Another topic that allows us to compare with prior findings is *climate change*. Previously, Feinberg and Willer (2013) showed that, on social media and news outlets, the moral rhetoric surrounding the environmental discourse primarily focuses on the *harm/care* foundation. While this may be true, we do not find significant evidence that an appeal to *care* signifies a positive or negative attitude toward climate change (OR=2.78, 95% CI=[0.64, 12.07]). This could be due to the fact that *care* is salient in arguments on both sides of the discourse, but each side portrays different framing patterns around the foundation, which is an interesting topic of examination for future work. Nevertheless, our results show that *fairness* is highly associated with the negative stance toward this topic (OR=0.10, 95% CI=[0.03, 0.38]). Tweets against climate change often focus on the unfair treatment of those who hold "alternative" viewpoints, such as in "Climate deniers is a term used to silence those pointing out the hypocrisy in the fanatical zeal on #climatetruth."

7 Conclusion

In this paper, we examine tools for characterizing moral foundations in social media content. We show that empirical findings based on MFT are specifically dependent on the scoring methods with which researchers label their data. Furthermore, we find that these methods, especially word count programs, often perform poorly and are biased in several ways. We instead propose Mformer, a language model fine-tuned on diverse datasets to recognize moral foundations, as an alternative classifier. We highlight the superior performance of Mformer compared to existing methods across several benchmarks spanning different domains. Using Mformer to analyze two datasets on Reddit and Twitter, we demonstrate its utility in detecting important patterns of moral rhetoric, such as conflicting judgments for the same moral dilemma that depend on the specific foundations upon which participants rely.

Limitations Labeling moral foundations is an inherently subjective task. We acknowledge that the scope of morality goes beyond what is observable on social media, and internet samples may not be representative of moral life as a whole. However, we think that social media datasets are large enough in size that one is able to draw important conclusions about how certain communities describe their moral concerns and judgments. Potential misuse of the proposed model and method could include content targeting and spreading mis-information.

Ethical considerations All datasets used in this paper are publicly available from prior work. In all analyses, we ensure that personal or potentially self-identifying information, such as usernames or URLs, is removed. The findings we present are descriptive insofar as morality is relevant to social issues debated online and we do not make normative claims within any of the examined domains.

Broader perspectives We primarily focus on the MFT because it is popular, which is in turn due to the availability of large annotated datasets. Even within this framework, the focus on how exactly each moral foundation is portrayed (e.g., as a vice or a virtue) is potentially important and can yield more fine-grained, novel results. In addition, alternative categorizations of moral beliefs based on competing theories exist and are worthy of examination. Among these, morality-as-cooperation is a rising candidate (Curry 2016; Curry, Mullins, and Whitehouse 2019), providing another set of dimensions with a different theoretical foundation.

Recognizing the prevalence of word count methods in detecting moral foundations, we believe a thorough evaluation

is warranted. This work establishes that, similar to other contexts such as sentiment analysis, lexicon-based word count programs often ignore contextual information, do not generalize well due to domain dependency, and may reveal social biases due to the way words are hand-chosen. Through a careful treatment of Mformer from training to (cross-domain) evaluation, we show that machine learning-based methods have potential to overcome these challenges.

Social media is a prolific resource for studying many aspects of morality, such as what moral dimensions are emphasized on both sides of a controversial issue. Findings from these studies can inform us about important social norms that guide debate on these issues, and can have practical implications for automated content moderation within online discussion forums as well as for the understanding of moral conflicts by machines. This work aims to caution researchers interested in this direction about the limitations of the available tools for measuring moral dimensions, and provide a more robust and reproducible alternative that has been evaluated across a range of benchmarks.

Acknowledgements

This work is supported by the CSIRO CRP Program and the Australian Research Council Project DP190101507. The authors would like to thank members of the Humanising Machine Intelligence Project at ANU for their feedback.

References

Amin, A. B.; Bednarczyk, R. A.; Ray, C. E.; Melchiori, K. J.; Graham, J.; Huntsinger, J. R.; and Omer, S. B. 2017. Association of Moral Values with Vaccine Hesitancy. *Nature Human Behaviour*, 1(12): 873–880.

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2022. Morality Beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures. *PsyArXiv*.

Botzer, N.; Gu, S.; and Weninger, T. 2022. Analysis of Moral Judgment on Reddit. *IEEE Transactions on Computational Social Systems*, 1–11.

Clifford, S.; Iyengar, V.; Cabeza, R.; and Sinnott-Armstrong, W. 2015. Moral Foundations Vignettes: A Standardized Stimulus Database of Scenarios Based on Moral Foundations Theory. *Behavior Research Methods*, 47(4): 1178–1198.

Clifford, S.; and Jerit, J. 2013. How Words Do the Work of Politics: Moral Foundations Theory and the Debate over Stem Cell Research. *The Journal of Politics*, 75(3): 659–671.

Curry, O. S. 2016. Morality as Cooperation: A Problem-Centred Approach. In Shackelford, T. K.; and Hansen, R. D., eds., *The Evolution of Morality*, 27–51. Cham: Springer International Publishing.

Curry, O. S.; Mullins, D. A.; and Whitehouse, H. 2019. Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies. *Current Anthropology*, 60(1): 47–69.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Fawcett, T. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8): 861–874.

Feinberg, M.; and Willer, R. 2013. The Moral Roots of Environmental Attitudes. *Psychological Science*, 24(1): 56–62.

Fleiss, J. L. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological bulletin*, 76(5): 378.

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.

Frimer, J. A.; Haidt, J.; Graham, J.; Dehgani, M.; and Boghrati, R. 2019. Moral Foundations Dictionary 2.0.

Garten, J.; Hoover, J.; Johnson, K. M.; Boghrati, R.; Iskiwitch, C.; and Dehghani, M. 2018. Dictionaries and Distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis. *Behavior Research Methods*, 50(1): 344–361.

González-Bailón, S.; and Paltoglou, G. 2015. Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 95–107.

Graham, J.; and Haidt, J. 2012. The Moral Foundations Dictionary.

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.

Guo, S.; Mokhberian, N.; and Lerman, K. 2023. A Data Fusion Framework for Multi-Domain Morality Learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 281–291.

Habernal, I.; and Gurevych, I. 2016. What Makes a Convincing Argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223. Austin, Texas: Association for Computational Linguistics.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

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.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.; Fisher, J.; and Weber, R. 2020. A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts. *Media and Communication*, 8(3): 164–179.

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.

Kobbe, J.; Rehbein, I.; Hulpuș, I.; and Stuckenschmidt, H. 2020. Exploring Morality in Argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, 30–40.

Koleva, S. P.; Graham, J.; Iyer, R.; Ditto, P. H.; and Haidt, J. 2012. Tracing the Threads: How Five Moral Concerns (Especially Purity) Help Explain Culture War Attitudes. *Journal of Research in Personality*, 46(2): 184–194.

Leetaru, K.; and Schrodt, P. A. 2013. GDELT: Global Data on Events, Location and Tone, 1979–2012. *ISA Annual Convention*, 2(4): 1–49.

Liscio, E.; Dondera, A.; Geadau, A.; Jonker, C.; and Murukannaiah, P. 2022. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2727–2745. Seattle, United States: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

McAdams, D. P.; Albaugh, M.; Farber, E.; Daniels, J.; Logan, R. L.; and Olson, B. 2008. Family Metaphors and Moral Intuitions: How Conservatives and Liberals Narrate Their Lives. *Journal of personality and social psychology*, 95(4): 978.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3): 1–23.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral Framing and Ideological Bias of News. In *Social Informatics*, volume 12467, 206–219. Cham: Springer International Publishing.

Nguyen, T. D.; Lyall, G.; Tran, A.; Shin, M.; Carroll, N. G.; Klein, C.; and Xie, L. 2022. Mapping Topics in 100,000 Real-Life Moral Dilemmas. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 699–710.

Pang, B.; and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in information retrieval*, 2(1–2): 1–135.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Puschmann, C.; and Powell, A. 2018. Turning Words into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media+ Society*, 4(3): 2056305118797724.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rezapour, R.; Dinh, L.; and Diesner, J. 2021. Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 177–188. Virtual Event USA: ACM.

Rezapour, R.; Shah, S. H.; and Diesner, J. 2019. 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*, 35–45. Minneapolis, USA: Association for Computational Linguistics.

Schwartz, H. A.; and Ungar, L. H. 2015. Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The Annals of the American Academy of Political and Social Science*, 659(1): 78–94.

Sechidis, K.; Tsoumakas, G.; and Vlahavas, I. 2011. On the Stratification of Multi-label Data. *Machine Learning and Knowledge Discovery in Databases*, 145–158.

Sim, J.; and Wright, C. C. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical therapy*, 85(3): 257–268.

Szymański, P.; and Kajdanowicz, T. 2017. A Network Perspective on Stratification of Multi-Label Data. In Torgo, L.; Krawczyk, B.; Branco, P.; and Moniz, N., eds., *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, 22–35. ECML-PKDD, Skopje, Macedonia: PMLR.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1): 24–54.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazizian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M. 2022. The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*.

van Leeuwen, F.; and Park, J. H. 2009. Perceptions of Social Dangers, Moral Foundations, and Political Orientation. *Personality and Individual Differences*, 47(3): 169–173.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187. Valencia, Spain: Association for Computational Linguistics.

Weinzierl, M. A.; and Harabagiu, S. M. 2022. From Hesitancy Framings to Vaccine Hesitancy Profiles: A Journey of Stance, Ontological Commitments and Moral Foundations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 1087–1097.

Zhou, K.; Smith, A.; and Lee, L. 2021. Assessing Cognitive Linguistic Influences in the Assignment of Blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 61–69. Association for Computational Linguistics.

Ziems, C.; Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3755–3773. Dublin, Ireland: Association for Computational Linguistics.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. This work introduces a robust, highly accurate tool for measuring moral foundations in text, which can help understand moral sentiment across different social contexts. To the best of our knowledge, the findings are descriptive and do not violate any of the above.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section 4 in the main text and Appendices D and E for more detail on our methodology.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes. We have discussed these in Section 7.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. Mformer and the findings in this paper are useful in characterizing moral sentiment on social media. We believe they do not pose any direct negative societal impacts.
   (g) Did you discuss any potential misuse of your work? Yes. This was discussed in Section 7 under "Limitations".
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. In Section 7.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, the hypothesis tests used in Sections 6.2 and 6.3 are described in more detail in Appendix H.2 and Appendix I.3, respectively.
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes, MFT is a prominent theory in moral psychology, as described in Section 2.1.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes. The results based on Mformer's predictions may challenge the validity of the claims by prior work, as we discussed in Section 6.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The main paper and supplemental material describe the dataset and training instructions in detail. Upon publication, these will be released publicly.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Section 4.2 in the main text and Appendix E.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Appendix E.2.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, the justification for the AUC metric is found in Section 4.3.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes. See Section 4.1.
   (b) Did you mention the license of the assets? Yes. All datasets are publicly available.
   (c) Did you include any new assets in the supplemental material or as a URL? No.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. For the Reddit dataset used in Section 6.1, we follow the process of removing URLs and potential self-identifying information by its curators Nguyen et al. (2022).
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA