Social Psychological and
Personality Science
2020, Vol. 11(8) 1057-1071
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1948550619876629
journals.sagepub.com/home/spp
SAGE

Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment

Joe Hoover[1], Gwenyth Portillo-Wightman[1], Leigh Yeh[1], Shreya Havaldar[1], Aida Mostafazadeh Davani[1], Ying Lin[2], Brendan Kennedy[1], Mohammad Atari[1], Zahra Kamel[1], Madelyn Mendlen[1], Gabriela Moreno[1], Christina Park[1], Tingyee E. Chang[1], Jenna Chin[1], Christian Leong[1], Jun Yen Leung[1], Arineh Mirinjian[1], and Morteza Dehghani[1]

Abstract

Research has shown that accounting for moral sentiment in natural language can yield insight into a variety of on- and off-line phenomena such as message diffusion, protest dynamics, and social distancing. However, measuring moral sentiment in natural language is challenging, and the difficulty of this task is exacerbated by the limited availability of annotated data. To address this issue, we introduce the Moral Foundations Twitter Corpus, a collection of 35,108 tweets that have been curated from seven distinct domains of discourse and hand annotated by at least three trained annotators for 10 categories of moral sentiment. To facilitate investigations of annotator response dynamics, we also provide psychological and demographic metadata for each annotator. Finally, we report moral sentiment classification baselines for this corpus using a range of popular methodologies.

Keywords

sentiment, NLP, text analysis, morality, Moral Foundations Theory

In this work, we introduce the Moral Foundations Twitter Corpus (MFTC), a collection of 35,108 tweets that have been hand annotated for 10 categories of moral sentiment. To facilitate the use of this corpus for theoretical and methodological research, we also provide baseline results for a wide range of models trained to detect moral sentiment in tweets. The motivation behind this work is to advance research at the intersection of psychology and natural language processing (NLP), an area that has received increasingly widespread attention in recent years. However, while a large portion of such research has focused on the task of inferring latent person-level traits and states (Iliev, Dehghani, & Sagi, 2014; Kern et al., 2016), such as personality (Azucar, Marengo, & Settanni, 2018; Garcia & Sikström, 2014; Park, Schwartz, & Eichstaedt, 2014), values (Boyd et al., 2015), and depression (Eichstaedt et al., 2018; Resnik, Garron, & Resnik, 2013; Zhou et al., 2015), this work is oriented toward a different task: measuring psychologically relevant constructs at the document level.

This task shares many similarities with standard sentiment classification tasks that focus on determining whether a "text", such as a tweet, expresses a particular sentiment such as positive or negative affect (for an accessible discussion of text analysis methods in psychology, see Iliev et al., 2014). However, it also introduces notable challenges such as the fact that moral sentiment categories co-occur, moral sentiment is often only implicitly signaled, and ground truth is, by definition, subjective. Despite these difficulties, research suggests that accounting for expressions of moral sentiment can afford insight into important downstream phenomena (Hoover, Dehghani, Johnson, Iliev, & Graham, 2017; Sagi & Dehghani, 2014) such as violent protest (Mooijman, Hoover, Lin, Ji, & Dehghani, 2018), charitable donation (Hoover, Johnson, Boghrati, Graham, & Dehghani, 2018), social avoidance (Dehghani et al., 2016), diffusion (Brady, Wills, Jost, Tucker, & Van Bavel, 2017), and political discourse (Dehghani, Sagae, Sachdeva, & Gratch, 2014; Johnson & Goldwasser, 2018).

However, a major obstacle for both theoretical and methodological research in this domain has been the difficulty of obtaining sufficient data. In our experience, all categories of moral sentiment have low base rates, which complicate

[1] University of Southern California, Los Angeles, CA, USA
[2] Rensselaer Polytechnic Institute, Troy, NY, USA

Corresponding Author:
Morteza Dehghani, University of Southern California, 3620 S. McClintock Ave., Los Angeles, CA 90089, USA.
Email: mdehghan@usc.edu

assembling a suitable corpus for annotation. Further, compared to sentiment domains like positive and negative valence or the basic emotions, annotating expressions of moral sentiment requires considerable domain expertise and training. Accordingly, conducting either theoretical or methodological research in this area has required substantial initial costs.

To address this issue, we have assembled a collection 35,108 tweets drawn from corpora focused around seven distinct, socially relevant discourse topics: All Lives Matter (ALM), Black Lives Matter (BLM), the Baltimore protests, the 2016 Presidential election, hate speech and offensive language (Davidson, Warmsley, Macy, & Weber, 2017), Hurricane Sandy, and #MeToo. Already, portions of this corpus have facilitated advances in both theoretical and methodological research. For example, Hoover, Johnson, Boghrati, Graham, and Dehghani (2018) relies on the Hurricane Sandy annotations to investigate the relationship between charitable donation and moral framing, and Mooijman, Hoover, Lin, Ji, and Dehghani (2018) use the Baltimore Protest annotations to predict violent protest from online moral rhetoric. These annotation sets have also been used for recent work advancing methods for measuring sentiment in natural language (Garten, Boghrati, Hoover, Johnson, & Dehghani, 2016; Garten et al., 2018; Lin et al., 2018).

While the limited availability of data has been a major obstacle for research in this area, the general absence of measurement baselines has also been a problem. As in any other area of psychological research, understanding the validity and relative performance of different approaches to measurement is essential for conducting reliable research and improving on current methodologies. Accordingly, we also report baseline results for multiple computational approaches to measuring moral sentiment in text. In addition to providing novel information about the relative performance of popular approaches to measuring moral sentiment in text, these baselines can inform future methodological innovation and help calibrate measurements of moral sentiment in other corpora.

Finally, we also provide psychological and demographic metadata for our annotators in order to facilitate investigations into annotator response patterns. In our view, accounting for annotator backgrounds is an important area for future research on sentiment analysis, particularly in domains characterized by high subjectivity such as moral values (Garten, Kennedy, Hoover, Sagae, & Dehghani, 2019; Garten, Kennedy, Sagae, & Dehghani, 2019). While, for example, an annotator's political ideology might not have a substantial influence on how they annotate "positive" and "negative" sentiment in a corpus of restaurant reviews, it seems likely that their ideology could substantially influence how they annotate expressions of moral values in a politically relevant corpus. We believe that developing a better understanding of these dynamics will be important as this area of research continues to develop. Accordingly, for each annotator, we provide responses to a range of psychological and demographic measures that can be used for investigations of annotator response patterns.

Our hope is that making these resources available for the research community will facilitate both theoretical and methodological advances by lowering the cost of conducting research in this area. Researchers can use these annotated tweets to evaluate new methods and train models for downstream application as well as to work on current problems in NLP, such as domain transfer and multitask learning (for discussion, see Ruder, 2017). To this end, we next provide a detailed description of the corpus, our annotation procedures, and a set of baseline classification results from a range of methods.

Corpus Overview

As noted above, the MFTC consists of 35,108 tweets drawn from seven different discourse domains. These domains were chosen for several reasons. First, we chose discourse domains related to issues that we know a priori are morally relevant in order to maximize the likelihood of selecting tweets that contain moral sentiment. Further, while many domains may seem to satisfy the constraint of being morally relevant, it was also necessary to select domains with sufficient popularity among Twitter users as, otherwise, we would not be able to obtain a sufficiently large sample of tweets.

Given these constraints, we strove to select a set of domains (1) that were relevant to current problems in the social sciences (e.g., prejudice, political polarization, natural disaster dynamics) and (2) that we expected a priori to contain a wide variety of moral concerns. Regarding the latter aim, we sought to accomplish this by selecting domains that were a priori associated with the political Left (e.g., BLM) or Right (ALM), both ideological poles (e.g., the Presidential election), or not aligned with either ideological group (e.g., Hurricane Sandy). Through these considerations, our goal was to maximize the variance in expressions of moral sentiment in the annotation corpus. This is particularly important, as the content of moral sentiment expressions can vary substantially with discourse context. For example, the moral sentiment contained in the BLM corpus is substantively distinct from the moral sentiment expressed in the Hurricane Sandy corpus, as these corpora focus on largely distinct issues. This heterogeneity makes out-of-domain prediction particularly difficult because expressions of moral sentiment in one domain will not necessarily generalize well to data drawn from a different domain. Accordingly, to help address this issue, we provide moral sentiment annotations for tweets drawn from multiple, heterogeneous contexts.

Annotation

Each tweet in the MFTC was labeled by at least three trained annotators (Total $N = 13$; see Table 1 for the distribution of annotators for each subcorpus) for 10 categories of moral sentiment as outlined in the Moral Foundations Coding Guide (see Appendix).

These categories are drawn from Moral Foundations Theory (Graham et al., 2013; Graham, Haidt, & Nosek, 2009), which proposes a five-factor taxonomy of human morality. In this

Table 1. Number of Tweets Annotated by *N* Annotators for Each Subdomain.

	Corpus						
N annotators	ALM	Baltimore	BLM	Election	Davidson	Sandy	#MeToo
3	4,316	4,496	28	659	4,959	4,591	2,522
4	108	575	388	4,699	2	—	2,006
5	—	522	4,837	—	—	—	62
6	—	—	—	—	—	—	295
7	—	—	—	—	—	—	5
8	—	—	—	—	—	—	1

Note. Cells show the number of tweets annotated by the number of annotators indicated under *N* annotators. ALM = All Lives Matter; BLM = Black Lives Matter.

model, each factor is bipolar, with each pole representing a virtue, or a prescriptive moral concern, and a vice, a prohibitive moral concern. The proposed factors (virtues/vices) are:

- *Care/harm*: prescriptive concerns related to caring for others and prohibitive concerns related to not harming others,
- *Fairness/cheating*: prescriptive concerns related to fairness and equality and prohibitive concerns related to not cheating or exploiting others,
- *Loyalty/betrayal*: prescriptive concerns related to prioritizing one's in-group and prohibitive concerns related to not betraying or abandoning one's in-group,
- *Authority/subversion*: prescriptive concerns related to submitting to authority and tradition and prohibitive concerns related to not subverting authority or tradition, and
- *Purity/degradation*: prescriptive concerns related to maintaining the purity of sacred entities, such as the body or a relic, and prohibitive concerns focused on the contamination of such entities.

While researchers often do not discriminate between the virtues and vices of a given foundation, their expressions in natural language are typically distinct and often independent. For example, an utterance focused on a harm violation (e.g., hurting someone emotionally or physically) is not necessarily also going to express care concerns. Accordingly, to account for the semantic independence between virtues and vices, each tweet in the corpus has been annotated for both.

Annotators, who were all undergraduate research assistants , participated in repeated training sessions during which they developed expert-level familiarity with the Moral Foundations Taxonomy. In early annotation stages, annotator disagreement was also addressed through discussion and, if necessary, subsequent label modification. However, moral sentiment is, in our view, qualitatively different from some other, more conventional, sentiment domains. In many cases, it is difficult to make a final determination of whether or not a document expresses moral sentiment or, for that matter, which moral sentiment it expresses, as such judgments are, ultimately, subjective (See Appendix for discussion).

Accordingly, while uniform annotator training is important, we believe that excessive focus on maximizing annotator agreement risks artificially inflating agreement at the cost of suppressing the natural variability of moral sentiment. Thus, while annotators were instructed to strive for consistency, they were also encouraged to avoid heuristics that might increase agreement with other annotators but would also lead them to neglect their own judgments.

Relying on this training, annotators were independently assigned to label each tweet from a subset of tweets sampled from a corpus associated with one of the seven discourse domains (see Table 2). The annotators used an annotation tool developed by Mooijman et al. (2018; this tool is available at https://github.com/limteng-rpi/moral_annotation_tool). Specifically, each tweet was assigned a label indicating the absence or presence of each virtue and vice or a label indicating that the tweet was nonmoral. This yielded a set of 11 labels for each tweet.

Annotator Metadata

For each annotator, we have also collected responses to a range of psychological and demographic measures. We provide measures of annotator's level of education, academic achievement (e.g., Scholastic Aptitude Test (SAT) score, Grade Point Average (GPA)), political ideology, political affiliation, Moral Foundations Values measured via the Moral Foundations Questionnaire (MFQ; Graham et al., 2009), analytic thinking (Toplak, West, & Stanovich, 2014), and everyday moral values (Hochreiter & Schmidhuber, 1997; Lovett, Jordan, & Wiltermuth, 2012). Annotators' Moral Foundations and political ideology scores were observed to skew liberal (see Table 3), which was expected due to the fact that annotators were drawn from the University of Southern California undergraduate student body.

These measures were obtained after the annotation process and thus were not used as criteria for selecting annotators. Further, while we have yet to fully incorporate these data into our own work, we suspect that accounting for and better understanding the association between annotators' individual differences and their annotations will be an important step for research in this domain.

General Sampling Procedure

To assemble the MFTC, we sampled tweets from larger corpora associated with each of the seven discourse domains (see Table 2). While, as noted above, these domains were selected

Table 2. Moral Foundations Twitter Corpus (MFTC) Discourse Domains.

Corpus	Corpus Description	Collection Method	Selection Criteria	*N*
All Lives Matter	Tweets related to the All Lives Matter movement	Purchased from Spinn3r.com	#AllLivesMatter, #BlueLivesMatter	4,424
Black Lives Matter (BLM)	Tweets related to the Black Lives Matter Movement	Purchased from Spinn3r.com	#BLM, #BlackLivesMatter	5,257
Baltimore Protests	Tweets posted during the Baltimore protests against the death of Freddie Gray	Purchased from Gnip.com	All tweets from cities with Freddie Gray protests	5,593
2016 U.S. Presidential Election	Tweets posted during the 2016 U.S. Presidential Election	Scraped via Twitter Application Programming Interface (API)	Followers of @HillaryClinton, @realDonaldTrump, @NYTimes, @washingtonpost, and @WSJ	5,358
Hurricane Sandy	Tweets related to Hurricane Sandy, a hurricane that caused record damage in the United States	Purchased from Gnip.com	#HurricaneSandy, #Sandy	4,591
#MeToo	Tweets related to the #MeToo movement	Purchased from Gnip.com	Random subset from 12 million tweets mentioning user IDs associated with high-profile allegations of sexual misconduct	4,891
Davidson Hate Speech	Tweets collected by Davidson et al. (2017) for hate speech and offensive language research	Obtained from Davidson et al. (2017)	Random sample from 85.4 million tweets that contained words in Davidson et al. (2017, Lexicon	4,873

Note. Metadata for each subcorpus contained in the MFTC. Subcorpora were collected via multiple methods, during varying time spans, and from distinct discourse domains. #BLM refers to Black Lives Matter. @WSJ is the official Twitter account for the *Wall Street Journal*.

Table 3. Annotator Moral Values and Political Ideology.

Moral Foundations			Political Ideology	
	Mean	*SD*		*N*
Care	3.67	.70	Very liberal	2
Fairness	3.55	.66	Liberal	5
Authority	1.96	.75	Slightly liberal	3
Loyalty	1.86	.76	Moderate	1
Purity	1.46	.86	Slightly conservative	1
—	—	—	Conservative	1
—	—	—	Very conservative	0

Note. Annotator metadata for the 13 Moral Foundations Twitter Corpus annotators. Moral Foundations measured on 0–5 scale. *SD* = standard deviation.

to maximize the base rates of moral sentiment, the proportion of tweets containing moral sentiment within each domain was still too low to use fully randomized sampling. Accordingly, our general sampling procedure relied on a combination of random sampling and semisupervised selection as in Garten et al. (2018) and Hoover et al. (2018).

Specifically, for each discourse domain, we used distributed dictionary representation (DDR; Garten et al., 2018) to calculate moral loadings for each tweet for each of the 10 virtues and vices. Then, for each virtue and vice, the 500 tweets with the highest loadings were selected for annotation. Finally, an additional 500 tweets were sampled from the subset of tweets with loadings that were ± 1 *SD* from 0.

This procedure yielded approximately 500 × 11 = 5,500 tweets per discourse domain. However, because virtues and vices regularly co-occur, some duplication is expected under this sampling procedure. Accordingly, as duplicates are removed, the final sampled *N* is less than the upper bound of 5,500.

Annotation Results

Overall, this annotation and sampling procedure yielded 4,000–6,000 annotated tweets for each discourse domain (see Table 4). It should be noted that the frequencies in Table 2 have been calculated based on annotators' majority vote, which was operationalized as receiving at least 50% agreement on the presence of a moral label. For example, if a particular tweet was annotated as "purity" by two of the four annotators, then that tweet would be marked as a positive case for purity concerns (see Table 5 for the distribution of majority vote moral labels that were decided by tie). It should also be noted that a particular tweet can be annotated for multiple labels based on this procedure.

Notably, the rates of each of the virtues and vices varies substantially across domain. For example, only approximately 2% of the ALM data (Total = 4,424) were labeled as degradation; while, in contrast, approximately 14% of the Sandy data (Total = 4,591) were labeled as degradation. These domain-level variations highlight the fact that the relevance of a particular moral concern to a given domain depends on the domain's content.

To evaluate interannotator agreement, we calculated both Fleiss's (1971) kappa and prevalence- and bias-adjusted Fleiss's kappa (PABAK; Sim & Wright, 2005) for multiple annotators (See Table 6). Fleiss's kappa represents the degree of observed agreement among annotators beyond what is

Table 4. Frequency of Tweets per Foundation Calculated Based on Annotators' Majority Vote.

Foundation	ALM	Baltimore	BLM	Election	Davidson	Sandy	#MeToo
Subversion	91	257	303	165	13	0	874
Authority	244	17	276	169	29	451	415
Cheating	505	519	876	620	80	434	685
Fairness	515	133	522	560	10	458	391
Harm	735	244	1037	588	195	179	433
Care	456	171	321	398	11	790	206
Betrayal	40	621	169	128	52	971	366
Loyalty	244	373	523	207	47	145	322
Purity	81	40	108	409	6	93	173
Degradation	122	28	186	138	106	636	941
Nonmoral	1,744	3,848	1,583	2,502	4,452	895	1,618
Total	4,424	5,593	5,257	5,358	4,961	4,591	4,891

Note. All tweets were annotated by at least three annotators. Majority vote was defined as $\geq$ 50% of annotators. ALM = All Lives Matter; BLM = Black Lives Matter.

Table 5. *N* and % Ties for Majority Vote Moral Labels Assigned by Even Number of Annotators.

	ALM	Baltimore	BLM	Election	Davidson	Sandy	#MeToo
N labels	162	311	349	2,979	1	0	2,533
% Ties	52.2	70.1	53.6	60.7	100	—	62.7

Note. Values in the first row indicate the total number of labels assigned by an even number of annotators. Values in the second row indicate the percentage of majority vote moral labels assigned by 50% of an even number of annotators (i.e., that were tied). ALM = All Lives Matter; BLM = Black Lives Matter.

expected by chance. However, it is strongly influenced by the prevalence of positive cases, and it can be difficult to interpret when applied to annotation data with skewed distributions of positive cases such as ours. PABAK adjusts for this (for discussion, see Sim & Wright, 2005) and offers an indication of the degree to which kappa is influenced by issues of prevalence or bias. As expected, due to the *b* of moral content across all corpora, all kappas were relatively low. However, adjusting for prevalence and bias suggests that interannotator agreement for each virtue and vice is reasonably high across discourse domains.

Baseline Computational Measurements of Moral Sentiment

While human annotation remains the most accurate method for measuring moral sentiment in text, the large sample sizes often used to investigate text-based moral sentiment usually necessitate supplementing human annotations with computational approaches. Such approaches range from word count methods, which rely on tallies of construct-relevant words to measure the presentence of a semantic construct, to machine learning pipelines that rely on state-of-the-art neural network architectures. Although various combinations of these methods have been used to investigate moral sentiment in text, there has been very little systematic investigation of their relative performance—that is, the degree to which they can reliably detect expressions of moral sentiment in natural language.

Accordingly, we next report classification baselines for a range of computational methods that have been used to measure moral sentiment in text. Specifically, we evaluate the degree to which five different approaches to measuring moral sentiment in text are able to identify MFTC messages that express moral sentiment, which we operationalize as messages that received a positive majority vote from human annotators. For each approach, we attempt to predict the document-level presence of moral sentiment for each of the five Moral Foundations both within and across each of the discourse domains represented in the MFTC. The performance baselines obtained through this experiment can serve as benchmarks for researchers investigating moral sentiment in other corpora, goals for researchers working on developing new methodologies for detecting moral sentiment in text, and guidelines for researchers trying to determine which methodological approach to use for a particular use-case.

Method

In order to provide a full-spectrum classification baseline for this corpus, we selected methodologies from a range of widely used approaches to sentiment classification. Specifically, we report results from four approaches. The first three approaches involve two steps: first, extracting "features" (e.g., word frequencies) from each tweet and then, second, using these features to train a classifier to predict whether a given tweet contains moral sentiment as indicated by human annotation majority vote. In this work, we use a support vector machine (SVM; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; James, Witten, Hastie, & Tibshirani, 2013) classifier for the classification step. In the fourth approach, we rely on a neural network classifier. In contrast to the other approaches, the neural network classifier is applied directly to each tweet, and, through an iterative optimization process, it learns which features predict moral sentiment.

Model Set 1. In the first of approach, we use the Moral Foundations Dictionary (MFD; Graham et al., 2009; available at

Table 6. Interannotator Agreement (PABAK and Kappa) Scores for All Datasets and Foundations.

		All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
All foundations	Kappa	.27	.16	.37	.38	.29	.19	.21	.27
	PABAK	.29	.20	.48	.41	.40	.52	.23	.29
Subversion	Kappa	.24	.19	.05	.53	.23	.08	.17	.24
	PABAK	.67	.88	.62	.89	.90	.96	.47	.67
Authority	Kappa	.29	.31	−.01	.54	.18	−.10	.19	.29
	PABAK	.71	.83	.85	.90	.89	.57	.67	.71
Cheating	Kappa	.42	.25	.27	.49	.41	.16	.36	.42
	PABAK	.75	.65	.70	.73	.79	.88	.68	.75
Fairness	Kappa	.33	.31	.17	.53	.44	.03	.33	.33
	PABAK	.85	.67	.85	.80	.80	.94	.77	.85
Harm	Kappa	.46	.20	.18	.39	.30	.37	.35	.46
	PABAK	.65	.49	.77	.61	.76	.88	.77	.65
Care	Kappa	.44	.25	.33	.42	.32	.10	.28	.44
	PABAK	.63	.64	.88	.82	.80	.97	.85	.63
Betrayal	Kappa	.18	.06	.24	.36	.17	.13	.18	.18
	PABAK	.82	.88	.65	.90	.90	.90	.69	.82
Loyalty	Kappa	.20	.23	.32	.64	.22	.11	.33	.20
	PABAK	.62	.77	.77	.87	.84	.90	.80	.62
Purity	Kappa	.16	.20	.23	.28	.19	.10	.28	.16
	PABAK	.91	.91	.95	.91	.76	.98	.88	.91
Degradation	Kappa	.19	.19	.11	.27	.22	.07	.28	.19
	PABAK	.89	.87	.94	.88	.90	.80	.52	.89
Nonmoral	Kappa	.33	.02	.57	.32	.29	.21	.36	.33
	PABAK	.62	.16	0.58	0.42	0.29	0.34	0.49	0.62

Note. Fleiss' kappa and prevalence- and bias-adjusted kappa (PABAK) for all annotations. kappa is strongly influenced by sparsity, PABAK adjusts for this influence and provides an indicator of how strongly a corresponding kappa is driven by prevalence or bias (see discussion in Annotation Results).

https://www.moralfoundations.org/othermaterials), a set of a priori selected words associated with each virtue and vice, to obtain message-level frequencies for words associated with each virtue and vice. These word counts were then used to train separate linear SVM (Drucker et al., 1997; James et al., 2013) models with ridge regularization to predict the binary presence of each Moral Foundation according to the majority vote human annotations, collapsing across virtues and vices. Each SVM was trained with C, a regularization parameter, set to 1 (for an introduction to SVM models, see James et al., 2013).

Model Set 2. For the second model set, we replaced the MFD with the MFD2(Frimer, Boghrati, Haidt, Graham, & Dehghani, 2015, available at https://osf.io/ezn37/), an updated lexicon of words associated with each virtue and vice. Using word frequencies based on the MFD2, we generated predictions of moral sentiment using linear SVMs with the same implementation as for Model Set 1.

Model Set 3. For the third model set, we again trained linear SVMs to predict moral sentiment; however, rather than relying on word counts, we used DDR (see Garten et al., 2018) to calculate moral loadings for each message. We used the same seed words for DDR as the ones used in the second study of Garten et al. (2018). These loadings represent the estimated similarity between a given message and latent semantic representations of each foundation. These loadings were then used as features to train a third set of linear SVMs.

Model Set 4. For the fourth model, we implemented and trained a multitask long short-term memory (LSTM; for an informal introduction to LSTMs, see Olah, 2015) neural network (Collobert & Weston, 2008; Luong, Le, Sutskever, Vinyals, & Kaiser, 2015) to predict moral sentiment. LSTMs are particularly effective for document-level classification tasks, as they rely on a recurrent structure that yields latent representations of documents that encode long-term dependencies among words. Here, we use a multitask architecture that involves training a model to predict labels for multiple outcomes. Specifically, for each discourse domain, we trained a multitask model to predict the document-level presence of each Moral Foundation.

To establish performance baselines, we first collapsed tweet annotations by taking the majority vote for each Foundation, where majority was considered $\geq 50\%$. We use this approach because it is a well-known and straightforward method for aggregating human annotations; however, we also believe that applying more sensitive annotation aggregation methods (e.g., see Passonneau & Carpenter, 2014; Paun et al., 2018) to the MFTC will be a fruitful area for future research. We then trained each model type separately on each discourse domain to predict each Moral Foundation. Then, using the entire corpus, we trained each model type to predict each Moral Foundation (i.e., "All" corpus). Finally, we also collapsed across Moral Foundations and trained each model type—on each discourse domain and the entire corpus—to predict whether documents were moral or not moral. All models were trained with 10-fold cross-validation to mitigate overfitting and

Table 7. Model F1, Precision, and Recall Scores for Moral Sentiment Classification.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.61 (.01)	.60 (.04)	.51 (.03)	.67 (.02)	.56 (.03)	.14 (.03)	.60 (.04)	.56 (.03)
	Precision	.52 (.01)	.73 (.02)	.61 (.03)	.88 (.03)	.71 (.04)	.93 (.05)	.84 (.03)	.42 (.04)
	Recall	.75 (.01)	.51 (.04)	.44 (.04)	.54 (.03)	.46 (.03)	.08 (.02)	.46 (.04)	.85 (.02)
SVM-MFD2	F1	.66 (.01)	.62 (.02)	.57 (.02)	.69 (.02)	.60 (.03)	.13 (.04)	.69 (.02)	.70 (.02)
	Precision	.58 (.01)	.54 (.03)	.59 (.03)	.88 (.02)	.74 (.03)	.77 (.23)	.85 (.03)	.59 (.03)
	Recall	.75 (.01)	.74 (.02)	.54 (.04)	.57 (.03)	.51 (.04)	.07 (.02)	.57 (.03)	.86 (.02)
SVM-DDR	F1	.71 (.01)	.65 (.03)	.62 (.03)	.79 (.01)	.71 (.02)	.14 (.04)	.78 (.01)	.75 (.02)
	Precision	.70 (.01)	.72 (.02)	.54 (.03)	.89 (.02)	.71 (.03)	.46 (.10)	.84 (.01)	.71 (.02)
	Recall	.73 (.01)	.59 (.03)	.75 (.04)	.72 (.02)	.72 (.03)	.08 (.03)	.73 (.03)	.81 (.02)
LSTM	F1	.80 (.01)	.76 (.02)	.69 (.03)	.89 (.01)	.77 (.01)	.14 (.03)	.81 (.02)	.86 (.01)
	Precision	.81 (.01)	.77 (.03)	.81 (.03)	.86 (.02)	.78 (.04)	.49 (.14)	.78 (.04)	.97 (.01)
	Recall	.79 (.01)	.76 (.02)	.61 (.04)	.92 (.02)	.76 (.04)	.08 (.02)	.84 (.02)	.77 (.02)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

Table 8. Model F1, Precision, and Recall Scores for Care.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.51 (.02)	.54 (.04)	.23 (.04)	.59 (.03)	.52 (.05)	.06 (.02)	.53 (.04)	.54 (.03)
	Precision	.49 (.02)	.53 (.05)	.16 (.03)	.65 (.03)	.49 (.05)	.93 (.09)	.50 (.06)	.43 (.04)
	Recall	.53 (.03)	.55 (.04)	.40 (.06)	.54 (.04)	.55 (.06)	.03 (.01)	.57 (.06)	.72 (.04)
SVM-MFD2	F1	.56 (.02)	.59 (.03)	.25 (.06)	.64 (.03)	.56 (.05)	.06 (.02)	.53 (.06)	.69 (.03)
	Precision	.64 (.02)	.65 (.05)	.17 (.04)	.61 (.04)	.48 (.05)	.89 (.08)	.47 (.07)	.68 (.04)
	Recall	.49 (.02)	.55 (.03)	.53 (.10)	.68 (.04)	.67 (.06)	.03 (.01)	.63 (.05)	.70 (.04)
SVM-DDR	F1	.48 (.02)	.55 (.03)	.23 (.04)	.61 (.02)	.48 (.04)	.06 (.02)	.43 (.04)	.69 (.03)
	Precision	.69 (.02)	.46 (.03)	.13 (.03)	.52 (.03)	.36 (.03)	.48 (.15)	.31 (.03)	.75 (.03)
	Recall	.37 (.02)	.68 (.05)	.71 (.07)	.75 (.03)	.74 (.06)	.03 (.01)	.70 (.08)	.65 (.04)
LSTM	F1	.63 (.02)	.65 (.05)	.26 (.04)	.77 (.02)	.61 (.06)	.06 (.02)	.36 (.11)	.78 (.03)
	Precision	.81 (.03)	.80 (.05)	.76 (.06)	.86 (.02)	.78 (.04)	.64 (.18)	.69 (.07)	.81 (.03)
	Recall	.52 (.02)	.55 (.05)	.16 (.03)	.70 (.03)	.50 (.08)	.03 (.01)	.25 (.10)	.75 (.04)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

approximate out-of-sample performance. To compare model sets, we rely on three performance metrics: *precision*, *recall*, and *F1*. *Precision*, the number of true positives divided by the number of predicted positives, represents the proportion of predicted positive cases that actually are positive cases. In contrast, *recall*, the number of true positives divided by the number of true positives and false negatives, represents the proportion of positive cases that the classifier correctly identifies. Finally, *F1*, the harmonic mean of *precision* and *recall*, provides a balanced summary of a classifier's ability to precisely identify true positives while also maximizing the proportion of true positives that are identified.

Results

As expected, performance varied substantially across methodology, discourse domain, and prediction task. Further, our results suggest that in the context of different domains and prediction tasks, each methodology showed different strengths and weaknesses. For example, while predictions derived from the LSTM models almost always outperformed predictions derived from the other models in terms of F1 and Precision, DDR generally yielded higher recall compared to both the LSTM- and dictionary-based approaches (see Tables 7–12). Notably, the results from DDR and LSTM models trained to predict only the presence of general moral sentiment, as opposed to a specific foundation, also suggest that poor performance may be a function of sparsity. That is, when all moral sentiment labels are collapsed into a single class, and there are thus more positive training observations, performance improves and stabilizes across discourse domains.

Finally, in some cases, the dictionary-based approaches also largely outperformed DDR in terms of precision. Finally, our results suggest, while, on average, the MFD and MFD2 dictionaries yield comparable performance in terms of F1, performance differences, again, depend on discourse domain and Foundation. Further, across discourse domains and Foundations, the MFD2 appears to offer higher precision compared

Table 9. Model F1, Precision, and Recall Scores for Fairness.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.47 (.02)	.57 (.04)	.30 (.06)	.52 (.05)	.55 (.06)	.03 (.01)	.42 (.04)	.32 (.06)
	Precision	.35 (.02)	.72 (.05)	.26 (.05)	.81 (.03)	.81 (.04)	.92 (.11)	.57 (.06)	.56 (.27)
	Recall	.72 (.03)	.48 (.05)	.35 (.08)	.38 (.05)	.42 (.06)	.01 (.00)	.33 (.04)	.28 (.13)
SVM-MFD2	F1	.61 (.01)	.59 (.03)	.39 (.05)	.68 (.05)	.70 (.04)	.02 (.02)	.63 (.04)	.59 (.03)
	Precision	.59 (.02)	.56 (.05)	.29 (.04)	.80 (.05)	.71 (.04)	.18 (.25)	.71 (.06)	.59 (.04)
	Recall	.63 (.02)	.62 (.03)	.60 (.08)	.60 (.05)	.69 (.06)	.01 (.01)	.58 (.05)	.59 (.05)
SVM-DDR	F1	.62 (.01)	.70 (.04)	.40 (.03)	.81 (.02)	.69 (.03)	.02 (.01)	.63 (.03)	.54 (.04)
	Precision	.79 (.01)	.63 (.06)	.26 (.03)	.78 (.03)	.59 (.04)	.42 (.25)	.56 (.03)	.85 (.04)
	Recall	.51 (.02)	.79 (.03)	.78 (.06)	.85 (.03)	.84 (.04)	.01 (.01)	.72 (.06)	.40 (.04)
LSTM	F1	.70 (.01)	.75 (.04)	.43 (.04)	.88 (.02)	.75 (.03)	.02 (.02)	.55 (.07)	.10 (.06)
	Precision	.81 (.02)	.84 (.04)	.81 (.07)	.91 (.02)	.85 (.03)	.35 (.22)	.76 (.04)	.06 (.04)
	Recall	.61 (.02)	.68 (.04)	.30 (.04)	.86 (.03)	.68 (.06)	.01 (.01)	.43 (.09)	.87 (.19)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

Table 10. Model F1, Precision, and Recall Scores for Loyalty.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.40 (.02)	.32 (.07)	.41 (.04)	.61 (.05)	.33 (.07)	.05 (.07)	.54 (.06)	.35 (.03)
	Precision	.38 (.01)	.26 (.07)	.38 (.04)	.69 (.05)	.24 (.05)	.08 (.07)	.58 (.06)	.47 (.04)
	Recall	.42 (.03)	.45 (.12)	.44 (.05)	.55 (.06)	.54 (.11)	.05 (.06)	.52 (.06)	.28 (.03)
SVM-MFD2	F1	.41 (.02)	.40 (.06)	.43 (.04)	.68 (.05)	.33 (.06)	.05 (.03)	.53 (.06)	.34 (.03)
	Precision	.40 (.02)	.56 (.10)	.38 (.04)	.80 (.05)	.23 (.04)	.22 (.15)	.52 (.07)	.51 (.05)
	Recall	.42 (.02)	.32 (.05)	.51 (.06)	.60 (.05)	.58 (.10)	.03 (.02)	.55 (.07)	.26 (.03)
SVM-DDR	F1	.36 (.02)	.37 (.03)	.46 (.05)	.73 (.04)	.27 (.04)	.05 (.03)	.52 (.03)	.36 (.04)
	Precision	.66 (.01)	.25 (.03)	.34 (.04)	.62 (.06)	.17 (.03)	.53 (.23)	.41 (.04)	.73 (.06)
	Recall	.25 (.02)	.74 (.06)	.72 (.08)	.88 (.03)	.75 (.08)	.02 (.01)	.70 (.05)	.24 (.03)
LSTM	F1	.43 (.02)	.38 (.09)	.50 (.03)	.87 (.03)	.26 (.03)	.03 (.01)	.38 (.07)	.40 (.05)
	Precision	.77 (.03)	.69 (.09)	.77 (.04)	.92 (.03)	.71 (.05)	.75 (.18)	.73 (.07)	.71 (.07)
	Recall	.30 (.03)	.26 (.08)	.37 (.03)	.83 (.07)	.16 (.02)	.02 (.01)	.26 (.07)	.28 (.05)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

Table 11. Model F1, Precision, and Recall Scores for Authority.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.42 (.01)	.55 (.07)	.16 (.04)	.73 (.04)	.42 (.07)	.00 (.00)	.39 (.05)	.43 (.05)
	Precision	.48 (.02)	.43 (.07)	.10 (.03)	.63 (.05)	.31 (.06)	.13 (.27)	.39 (.06)	.41 (.05)
	Recall	.38 (.02)	.77 (.10)	.36 (.07)	.86 (.03)	.70 (.05)	.00 (.00)	.39 (.05)	.45 (.06)
SVM-MFD2	F1	.40 (.02)	.41 (.07)	.19 (.04)	.68 (.06)	.38 (.04)	.01 (.01)	.38 (.04)	.40 (.03)
	Precision	.53 (.03)	.74 (.08)	.12 (.03)	.57 (.07)	.26 (.04)	.58 (.38)	.39 (.04)	.69 (.04)
	Recall	.33 (.02)	.29 (.06)	.57 (.12)	.85 (.03)	.68 (.04)	.00 (.00)	.38 (.04)	.29 (.03)
SVM-DDR	F1	.36 (.02)	.48 (.05)	.19 (.03)	.73 (.04)	.30 (.04)	.01 (.01)	.47 (.03)	.50 (.04)
	Precision	.71 (.02)	.33 (.04)	.11 (.02)	.60 (.06)	.18 (.03)	.50 (.25)	.37 (.03)	.78 (.04)
	Recall	.24 (.02)	.85 (.06)	.73 (.08)	.93 (.04)	.81 (.06)	.01 (.00)	.65 (.05)	.37 (.04)
LSTM	F1	.47 (.02)	.57 (.07)	.19 (.02)	.83 (.03)	.33 (.07)	.01 (.01)	.47 (.03)	.59 (.03)
	Precision	.80 (.02)	.85 (.07)	.77 (.07)	.91 (.05)	.80 (.09)	.24 (.31)	.67 (.06)	.80 (.06)
	Recall	.34 (.02)	.43 (.07)	.11 (.01)	.76 (.06)	.21 (.06)	.01 (.01)	.36 (.03)	.46 (.03)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

Table 12. Model F1, Precision, and Recall Scores for Purity.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	.30 (.03)	.15 (.01)	.07 (.02)	.54 (.06)	.35 (.05)	.03 (.01)	.45 (.06)	.20 (.10)
	Precision	.43 (.04)	.08 (.01)	.04 (.01)	.47 (.08)	.29 (.06)	.96 (.06)	.47 (.06)	.38 (.17)
	Recall	.23 (.03)	.82 (.06)	.43 (.13)	.64 (.09)	.45 (.05)	.02 (.00)	.44 (.06)	.14 (.08)
SVM-MFD2	F1	.33 (.02)	.34 (.03)	.13 (.06)	.49 (.07)	.43 (.07)	.02 (.01)	.51 (.04)	.20 (.06)
	Precision	.59 (.03)	.73 (.07)	.07 (.04)	.36 (.06)	.33 (.06)	.30 (.35)	.50 (.04)	.64 (.16)
	Recall	.23 (.02)	.22 (.02)	.54 (.13)	.76 (.10)	.64 (.07)	.01 (.01)	.53 (.05)	.12 (.03)
SVM-DDR	F1	.24 (.02)	.25 (.04)	.11 (.03)	.34 (.07)	.33 (.05)	.03 (.01)	.54 (.05)	.14 (.03)
	Precision	.66 (.03)	.15 (.03)	.06 (.02)	.21 (.05)	.22 (.04)	.35 (.16)	.45 (.04)	.74 (.13)
	Recall	.15 (.02)	.76 (.12)	.88 (.09)	.84 (.09)	.72 (.07)	.01 (.01)	.69 (.06)	.08 (.02)
LSTM	F1	.41 (.02)	.57 (.07)	.07 (.03)	.48 (.10)	.47 (.05)	.04 (.02)	.53 (.07)	.15 (.03)
	Precision	.80 (.03)	.85 (.07)	.81 (.24)	.81 (.10)	.79 (.08)	.48 (.19)	.71 (.08)	.72 (.10)
	Recall	.28 (.02)	.43 (.07)	.03 (.02)	.34 (.10)	.33 (.04)	.02 (.01)	.43 (.07)	.09 (.02)

Note. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate *SD*s across folds. ALM = All Lives Matter; BLM = Black Lives Matter; SVM = support vector machine; MFD = Moral Foundations Dictionary; DDR = distributed dictionary representation; LSTM = long short-term memory; *SD* = standard deviation.

to the original MFD. In contrast, the original MFD appears to offer generally better recall compared to the MFD2.

Together, our classification results demonstrate the viability of measuring moral sentiment in natural language using a range of methodologies; however, they also highlight the difficulty of this task. Regardless of methodology, considerable performance variation was observed across both discourse domain and Foundation. In our view, this raises multiple important goals for future research such as working toward a better understanding of the causes of this variation and developing methodological approaches that minimize it.

Discussion

By understanding and measuring the expression of moral sentiment in natural language, researchers can gain insight into a variety of important digital- and real-world phenomena (Hoover, Johnson-Grey, et al., 2017; Sagi & Dehghani, 2014). However, in practice, it can be quite costly to take advantage of these opportunities. In our view, a major driver of this cost has been the difficulty of obtaining annotated data, which is necessary for evaluating method performance and training supervised language models.

To address this issue, we have developed the MFTC, a collection of 35,108 tweets drawn from seven different domains and annotated for 10 types of moral sentiment. Using the MFTC, we also report classification baselines for a range of approaches to measuring moral sentiment in text. Finally, we also report individual difference measures for each annotator, so that researchers can investigate the potential effects of annotator characteristics on the annotation process.

Researchers can use this corpus to train supervised models for predicting moral sentiment in new data. For example, researchers interested in measuring expressions of moral sentiment in a new sample of tweets collected from one of the MFTC domains could train a classifier on the MFTC and then use that classifier to predict moral sentiment in the new sample. Alternatively, researchers could also use a MFTC-trained classifier to predict moral sentiment in a new sample taken from a different domain of discourse. However, for such applications, it is important to note that expressions of moral sentiment are often domain specific. For instance, the moral relevance of "Freddie Gray," the name of the Black man whose death in police custody triggered the Baltimore Protests, is likely very different in the BLM corpus compared to the ALM corpus. Accordingly, we would encourage researchers interested in measuring moral sentiment in domains not included in the MFTC to use the MFTC to supplement their own annotations. For instance, they could annotate a portion of tweets collected from the new domain and then combine these annotations with the MFTC to train a domain-specific classifier that is also informed by the MFTC annotations. In our view, this may be a particularly useful approach, as it equates to using the MFTC to mitigate the limiting issues of sparsity.

The MFTC can also facilitate new methodological research on computational measurement of moral sentiment. While our baseline results suggest that, in most cases, state-of-the-art approaches such as LSTMs outperform simpler approaches, these performance differences appear to vary substantially across discourse domains. Using the MFTC, researchers can develop a better understanding of what drives these variations, find ways to integrate the strengths of distinct methodological approaches, and, ultimately, develop methods that are able to more directly address the difficulties observed in moral sentiment classification.

Finally, relying on the annotator metadata included with the MFTC, researchers can begin investigating the effects that annotator individual differences may have on annotation outcomes. Developing a better understanding of these dynamics is particularly important for moral sentiment analysis, as moral sentiment is an inherently subjective construct. For instance, future research could focus on integrating approaches to representing "ground truth" that are more sophisticated than "majority vote," such as approaches based on cultural consensus theory (Romney, Weller, & Batchelder 1986; Weller &

Mann, 1997). To directly address issues of annotator characteristics, researchers could also use model-based approaches to measuring ground truth from human annotations (e.g., see Paun et al., 2018), which can be extended to include annotator characteristics. While such investigations likely would require additional annotations, our hope is that researchers will make them public and thus extensions of the MFTC. By adding to the MFTC over time, it could become an even more useful resource for investigating moral sentiment in natural language.

Open data standards regarding annotated text corpora are a key element in the emerging field of computational social science. They afford greater research transparency and can help facilitate scientific progress via the free dissemination of materials that are costly to assemble. Our hope is that, as more researchers use the MFTC, the resources we provide here will be continually expanded. Through the MFTC, our goal is to contribute to this culture of openness and thereby help facilitate both applied and methodological advances in the computational social sciences.

Appendix

Moral Foundations Coding Guide

Moral values play many important roles in human social functioning. They influence our judgments and behaviors (DeScioli & Kurzban, 2013; Ellemers, van der Toorn, Paunov, & van Leeuwen, 2019; Greene, 2014; Haidt, 2012) and help coordinate complex social dynamics (Curry, Jones Chesters, & Van Lissa, 2019; DeScioli & Kurzban, 2013; Enke, 2019). However, moral values are not always visible in day-to-day life. You cannot, for example, necessarily induce a person's moral values simply by considering their appearance or behavior. To make up for this invisibility, people often rely on language to signal their moral values (Keen, 2015; Poulshock, 2006; Smead, 2010). While such signals can be ingenuous or disingenuous, expressions of moral sentiment serve as informationally rich indicators of individuals' and groups' moral values or at least the moral values they wish to display.

Combined with the availability of large-scale natural language data, the tendency for people to encode moral values in their daily language presents a valuable opportunity for empirical research on the real-world function of moral values (Hoover, Dehghani et al., 2017). Here, to help facilitate such research, we report a method for annotating moral sentiment in natural language. Specifically, we introduce the theoretical framework that we rely on to operationalize moral values, Moral Foundations Theory (MFT; Graham et al., 2013, 2009), discuss our approach to annotating moral sentiment, and provide example instructions for annotators.

Expressions of Moral Values in Language

When people write or speak about phenomena that they moralize, they may employ a variety of rhetorical strategies to communicate their values (Keen, 2015). Often, these strategies rely on words that are explicitly normative, such as "right," "wrong," "good," or "bad"; however, in many cases, people's word choices do not just indicate moralization but also provide specific information about the moral relevance of a particular expression. For example, when a person describes an event they perceive to be unfair, they might say something like "I can't believe that happened. It's so unfair!" or, slightly less explicitly, "They had NO RIGHT to do that." Such language not only signals moralization but also provides information that can be used to make inferences about the specific nature of a particular expression of moral sentiment.

A Taxonomy of Moral Sentiment

To impose structure on the many possible forms of moral sentiment, we rely on MFT (Graham et al., 2013, 2009), a pluralistic, psychological model of moral values. MFT suggests that peoples' moral values cluster into five primary bipolar dimensions or *foundations* of moral concerns that represent the morally bad (vices) at one extreme and the morally good (virtues) at the other. As described on YourMorals.org (2019), the proposed foundations are:

1. *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.
2. *Fairness/cheating*: This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy.
3. *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."
4. *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.
5. *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 that can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

To the extent that linguistic cues signaling associations with different foundations can be identified, researchers can detect and quantify expressions of moral values by analyzing natural language artifacts (e.g., written documents and speech). In early work, Graham, Haidt, and Nosek (2009) demonstrated this by using frequencies of foundation-relevant words to measure differences in moral values sentiment between conservative and liberal sermons. More recently, researchers have shown that moral sentiment analysis based on the MFT taxonomy can fruitfully be applied to a range of applications and domains (Dehghani et al., 2016; Hoover et al., 2018; Mooijman et al., 2018; Sagi & Dehghani, 2014).

Annotating Moral Sentiment in Natural Language

We define the task of annotating moral sentiment in natural language as determining which, if any, categories of moral values are relevant to a given document. In our research, we rely on the taxonomy proposed by MFT to identify categories of moral values. However, even when relying on the MFT framework, researchers must make several initial decisions regarding their approach to annotating MFT values.

First, they must decide what dimensions of MFT to code for. If a foundation-specific hypothesis will be tested, it might make sense to code only for that foundation. Frequently, though, it is necessary to code for multiple foundations. In such cases, the most obvious approach is to code for the presence of each of the five foundations. However, some research programs might require more fine-grained labels. While the poles of each dimension are related, they also express distinct sentiments that might be psychologically relevant. For example, the statement "We must end suffering" is likely not psychologically equivalent to the statement "We must provide kindness and compassion." It can thus also be useful to code for the presence of each pole of each foundation, which yields 10 individual codes. Additionally, it is important to code nonmoral texts as such. Thus, in our work, an annotation procedure may require labeling each document for up to 11 categories.

Researchers must also decide how to handle overlapping labels (e.g., expressions of moral sentiment that are associated with multiple foundations). In our work, we always allow for overlapping labels during the annotation process. In some instances, we have also asked annotators to distinguish between the primary domain moral sentiment expressed in a document and potential secondary domains of moral sentiment. However, during reliability analyses, we found that while coders sufficiently agreed about whether a given moral sentiment was present, they agreed far less about which domain was most dominant. Accordingly, we suggest that researchers code for the presence/absence of each foundation.

Training Human Annotators

After selecting an annotation framework, we suggest that researchers develop a clear protocol for determining the moral domains relevant to a given document. In our experience, this is particularly important, as making this determination can be quite difficult in practice. This difficulty is largely driven by two sources of ambiguity. The first is ambiguity regarding the foundations that a moral expression is associated with. For example, an instance of moral sentiment could appear to be strongly associated with authority but also potentially associated with loyalty. In such cases, it is ambiguous whether the instance should be labeled as just authority or both authority and loyalty.

The second source of ambiguity is caused by the difficulty of inferring the moral relevance intended by an author and the moral relevance encoded in their language. For example, a social media message might simply state that the author thinks "Everything that is going on with abortion these days is reprehensible." In this case, it is clear that this is likely a morally relevant statement, but it is less clear what foundation this statement is relevant to. If we knew that the author was a secular liberal concerned about civil rights, we might assume that the author is concerned about violations of women's reproductive rights and therefore infer that this expression is most strongly associated with the fairness/cheating foundation. In contrast, if we knew that the author was a conservative Christian, we might assume that the author was expressing an anti-abortion sentiment, perhaps associated with purity/degradation. Thus, the same expression can be meant to convey divergent moral sentiments, and it is not infrequent that competing interpretations of a particular expression are difficult if not impossible to resolve systematically.

These ambiguities present considerable challenges for human annotators who must strike an acceptable balance between exploiting often weak signals of moral sentiment while also avoiding unfounded speculation about authorial intent. If each coder on a team relies too extensively on their own intuitions to infer the moral sentiment expressed in a document, coder reliability will almost certainly suffer. On the other hand, constraining interpretations of sentiment to a plane that is too literal can be equality problematic because the subtleties of human language and morality will likely be lost. For example, an explicit coding of the text, "OMG I am going to MURDER THAT DOG, it is just TOO CUTE" references an act that is definitively relevant to harm. However, taken in context, the profession of murderous intent is clearly idiomatic, and it is not at all clear that this statement should be coded as equivalent to a statement like, "I hate him. He deserves to be murdered." Thus, a careful balance must be struck between implicit coding—coding that relies on inferences made about authorial intent—and explicit coding—coding that relies exclusively on literal interpretations of textual content.

While this is difficult to do well and impossible to do perfectly, maintaining an awareness of these two extremes can hopefully limit coder biases in either direction. Because we typically do not have access to the authors of the texts we are analyzing—and sometimes not even to the discourse they are embedded in—we train annotators to limit the degree to which they base their annotations on what they think the author meant to express. That is, our approach to annotating moral sentiment in natural language focuses primarily on explicit signals of moral sentiment and minimizes reliance on inferred authorial intent unless inferences are strongly defensible. We take this approach in order to reduce the risk of annotator cultural biases introducing additional noise to the annotations. However, while our annotation protocol generally aims to reduce annotator disagreement, we also suggest avoiding artificial reduction of annotation variance between annotators.

When coding for MFT content, there are often disagreements about which foundation is relevant to a given statement, and in many of these cases, even among expert coders, it is not clear which perspective is correct. Of course, such disagreements can be resolved through discussion, and to an extent, resolution is appropriate. However, at some point, resolution of coder disagreement begins to artificially inflate intercoder reliability. Moral values are inherently subjective phenomena, and the true accuracy of a code cannot really be determined because there is no objective criterion to which it can be compared. The closest we can come to an objective criterion is consensus among some constituency. As consensus is approached, the certainty that a given phenomenon has a strong subjective association with a specific Moral Foundation increases. This means that low consensus—for example, among trained coders—is not simply a problem that must be resolved; it is a potential indication that the association between a foundation and a phenomenon might be subject to important boundary conditions, weak or even illusory. Training coders so that they disagree minimally does not change this fact but rather hides it. Consequently, while coders obviously need training, annotator training should focus on establishing a shared network of concepts and a few heuristics that can be used for generating codes but also that this training should stop short of fabricating agreement.

Instructions for Annotators

Each moral virtue and vice should be coded as the capitalized first letter of the foundation and a capitalized P if the foundation is a virtue and an *N* if the foundation is a vice. P and N correspond to Positive and Negative. If a document does not have any moral content, it should be coded as NM, which corresponds to nonmoral. The entire scheme is:

- *Care*: HP
- *Harm*: HN
- *Fairness*: FP
- *Cheating*: FN
- *Loyalty*: LP
- *Betrayal*: LN
- *Authority*: AP
- *Subversion*: AN
- *Purity*: PP
- *Degradation*: PN
- *Nonmoral*: NM

The documents that you will code will be displayed in a spreadsheet. Each document will be contained in a row. Your job is to add the above-described labels to the column called "mft_sentiment" (see Table A1 for an example).

Table A1. Spreadsheet Example.

ID	Text	mft_sentiment	Notes
123	Cause I want to be anarchy, Its the only way to be	AN	

When coding a document, try to follow this order of operations:

1. Does the document (or sentence/phrase) seem to have moral content? If no, enter "NM" in the "mft_sentiment" column. If yes, continue to 2.
2. Which foundation does the document seem most associated with? Label the document with this code in the "mft_sentiment" column and proceed to 3.
3. Does the document seem to be associated with any other foundations? If no, proceed to 4. If yes, add these to the "mft_sentiment" column.
4. Is there anything important to note about the document? Is it in a different language? Does it seem in any way like it should be excluded (e.g., because it is fake, because it has been repeated multiple times) If so, add a note describing these issues in the "note" column.

While the language that is used to signal moral relevance can be highly variable, abstract, and context dependent, it can be useful to have a core set of terms or concepts to use as anchors for identifying the presence or absence of a signal. Below such terms are listed in association with each foundation pole:

Harm

Positive (HP): save, defend, protect, compassion
Negative (HN): harm, war, kill, suffer

Fairness

Positive (FP): fair, equal, justice, honesty
Negative (FN): unfair, unequal, unjust, dishonest

Loyalty

Positive (LP): solidarity, nation, family, support
Negative (LN): betray (in-group), abandon, rebel (against in-group)

Authority

Positive (AP): duty, law, obligation, order
Negative (AN): rebel (against authority), chaos, disorder, betray (your role)

Purity

Positive (PP): sacred, preserve, pure, serenity
Negative (PN): dirty, repulsive, disgusting, revolting

In addition to coding documents, please keep a log of difficult cases that you encounter. We will discuss these with the other coders, and, when it is appropriate, we will try to resolve these cases.

For example, if prayer is mentioned in a document, this could be relevant to purity; however, if someone is praying for another person's well-being, it might also be relevant to care (HP). While it is certainly acceptable to code both as present, some degree of conservatism should be maintained because we do not want to inflate the frequency of a given moral sentiment.

Authors' Note

Gwenyth Portillo-Wightman, Leigh Yeh, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian contributed equally to this work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been funded in part by NSF IBSS #1520031, NSF CAREER BCS-1846531, and the Army Research Lab.

ORCID iD

Morteza Dehghani https://orcid.org/0000-0002-9478-4365

References

Azucar, D., Marengo, D., & Settanni, M. (2018, April). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. *Ninth International AAAI*. Palo Alto, CA: The AAAI Press.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017, July). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 7313–7318.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167).

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019, February). Mapping morality with a compass: Testing the theory of "morality-as-cooperation" with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*. Palo Alto, CA: AAAI Press.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., . . . Graham, J. (2016, January). Purity homophily in social networks. *Journal of Experimental Psychology General*, *145*, 366–375.

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, *11*, 1–14.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*, 477–496.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161). California: Neural Information Processing Systems, Inc.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., . . . Schwartz, H. A. (2018, October). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 11203–11208.

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019, January). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*. doi:10.1177/1088868318811759

Enke, B. (2019). Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, *134*, 953–1019.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378.

Frimer, J., Boghrati, R., Haidt, J., Graham, J., & Dehghani, M. (2019). Moral Foundations Dictionary 2.0. 11 April. https://doi.org/10.17605/OSF.IO/EZN37

Garcia, D., & Sikström, S. (2014, September). The dark side of Facebook: Semantic representations of status updates predict the dark triad of personality. *Personality and Individual Differences*, *67*, 92–96.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In S Kambhampati (ed.), *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*. New York, New York:

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

Garten, J., Kennedy, B., Hoover, J., Sagae, K., & Dehghani, M. (2019). Incorporating demographic embeddings into language understanding. *Cognitive Science*, *43*, e12701.

Garten, J., Kennedy, B., Sagae, K., & Dehghani, M. (2019). Measuring the importance of context when modeling language comprehension. *Behavior Research Methods*, *51*, 480–492.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. *47*, pp. 55–130). Amsterdam, the Netherlands: Elsevier.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Greene, J. (2014). *Moral tribes: Emotion, reason and the gap between us and them*. London, England: Atlantic Books.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Vintage.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoover, J., Dehghani, M., Johnson, K., Iliev, R., & Graham, J. (2017). Into the wild: Big data analytics in moral psychology. In J. Graham & K. Gray (Eds.), *The atlas of moral psychology*. New York, NY: Guilford Press.

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2018, April). Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, *4*, 9.

Iliev, R., Dehghani, M., & Sagi, E. (2014, July). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*, 265–290.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. *112*, p. 18). New York, NY: Springer.

Johnson, K., & Goldwasser, D. (2018). Classification of moral foundations in microblog political discourse. In I. Gurevych & Y Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. *1*, pp. 720–730). Melbourne, Australia: Association for Computational Linguistics.

Keen, I. (2015, December). The language of morality. *The Australian Journal of Anthropology*, *26*, 332–348.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016, December). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*, 507–525.

Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., & Ji, H. (2018). Acquiring background knowledge to improve moral value prediction. In J. Diesner, E Ferrari, & G Xu (eds.), *The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM2018)*. New York, NY: Association of Computing Machinery.

Lovett, B. J., Jordan, A. H., & Wiltermuth, S. S. (2012, July). Individual differences in the moralization of everyday life. *Ethics & Behavior*, *22*, 248–257.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. 14 November. *arXiv preprint* arXiv:1511.06114.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018, June). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*, 389–396.

Olah, C. (2015). Understanding LSTM networks. Retrieved August 8, 2019, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Park, G., Schwartz, H., & Eichstaedt, J. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*:934–952.

Passonneau, R. J., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, *2*, 311–326.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., & Poesio, M. (2018). Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, *6*, 571–585.

Poulshock, J. W. (2006). Language and morality: Evolution, altruism, and linguistic moral mechanisms (Unpublished doctoral dissertation). University of Edinburgh, United Kingdom.

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1348–1353). Stroudsburg, PA: Association for Computational Linguistics.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. Retrieved from http://arxiv.org/abs/1706.05098

Sagi, E., & Dehghani, M. (2014, April). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*, 132–144.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257–268.

Smead, R. (2010). Indirect reciprocity and the evolution of “moral signals”. *Biology & Philosophy*, *25*, 33–51.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014, April). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*, 147–168.

Weller, S. C., & Mann, N. C. (1997). Assessing rater performance without a “gold standard” using consensus theory. *Medical Decision Making*, *17*, 71–79.

YourMorals.org. (2019). Moralfoundations.org. Retrieved July 18, 2016, from http://moralfoundations.org/

Zhou, L., Baughman, A. W., Lei, V. J., Lai, K. H., Navathe, A. S., Chang, F., . . . Rocha, R. A. (2015). Identifying patients with depression using free-text clinical documents. *Studies in Health Technology and Informatics*, *216*, 629–633.

Author Biographies

Joe Hoover is a computational social psychologist and postdoctoral research fellow in the Kellogg School of Management at Northwestern University. He studies the psychological underpinnings of conflict and violence using methods drawn from the social and computational sciences.

Gwenyth Portillo-Wightman is an undergraduate psychology research assistant at the University of Southern California, studying cognitive science and computer science. Her interests include computational linguistics, machine learning, and artificial intelligence, especially for social good applications.

Leigh Yeh is a Master’s student in the Department of Computer Science at the University of Southern California with an undergraduate background in cognitive science from the University of Michigan. Her research interests include computational social science, computational linguistics, and cognitive science.

Shreya Havaldar is an undergraduate research assistant in the Department of Computer Science at the University of Southern California.

Aida Mostafazadeh Davani is a PhD student in the Department of Computer Science at the University of Southern California. She is interested in analyzing the bias in language by applying natural language processing methods.

Ying Lin is a PhD student in the Department of Computer Science at Rensselaer Polytechnic Institute.

Brendan Kennedy is a PhD student in the Department of Computer Science at the University of Southern California. His interests are in the intersection of psychology, computational linguistics, and machine learning.

Mohammad Atari is a doctoral student in social psychology at the University of Southern California, where he blends his interests in morality, culture, and evolution using computational methods. His ongoing work includes socio-ecological perspectives on understanding inter-group variation in moral values.

Zahra Kamel, **Madelyn Mendlen**, **Gabriela Moreno**, **Christina Park**, **Tingyee E. Chang**, **Jenna Chin**, **Christian Leong**, **Jun Yen Leung**, and **Arineh Mirinjian** are undergraduate research assistants in the Department of Psychology at the University of Southern California.

Morteza Dehghani is an assistant professor of psychology and computer science at the University of Southern California. His research, which spans the boundary between psychology and artificial intelligence, investigates the properties of cognition at multiple levels of analysis.

Handling Editor: Simine Vazire