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, Ying Lin[2], Davani[1], 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‡], Morteza Dehghani[1]

[1]University of Southern California

[2]Rensselaer Polytechnic Institute

Paper in Press at Social Psychological and Personality Science

Author Note

[*] Contributed equally. [†] Contributed equally. [‡]Contributed equally.

This work has been funded in part by NSF IBSS #1520031, NSF CAREER BCS-1846531, and the Army Research Lab. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Correspondence regarding this article should be addressed to Morteza Dehghani, mdehghan@usc.edu, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061.

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 meta-data 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

Moral Foundations Twitter Corpus:

A collection of 35k tweets annotated for moral sentiment

In this work, we introduce the Moral Foundations Twitter Corpus, a collection of 35,108 tweets that have been hand annotated for 10 categories of moral sentiment. To facilitate 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, 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, which 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 complicates 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, Black Lives Matter, the Baltimore protests, the 2016 Presidential election, hate speech & 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 et al. (2018) relies on the Hurricane Sandy annotations to investigate the relationship between charitable donation and moral framing, and Mooijman et al. (2018) uses 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 also inform future methodological innovation and help calibrate measurements of moral sentiment in other corpora.

Finally, we also provide psychological and demographic meta-data 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 work on current problems in natural language processing (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 Moral Foundations Twitter Corpus (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 Black Lives Matter 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 sub-corpus) for 10 categories of moral sentiment as outlined in the Moral Foundations Coding Guide (See Appendix A).

Table 1

Number of Tweets Annotated by N Annotators for each Sub-domain

	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.

These categories are drawn from Moral Foundations Theory (MFT; Graham et al., 2013; Graham, Haidt, & Nosek, 2009), which proposes a five-factor taxonomy of human morality. In this 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 ingroup and prohibitive concerns related to not betraying or abandoning one's ingroup.
- **Authority/Subversion**. Prescriptive concerns related to submitting to authority and tradition and prohibitive concerns related to not subverting authority or tradition.
- **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 (authors 8-16, and others), 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 (Hoover, Johnson-Grey, Dehghani, & Graham, 2017).

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 seven discourse domains (See Table 3). The annotators used an annotation tool developed for Mooijman et al. (2018)[1]. 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 non-moral. This yielded a set of 11 labels for each tweet.

Annotator Meta-data. For each annotator, we have also collected responses to a range of psychological and demographic measures. We provide measures of annotator level of education, academic achievement (e.g. SAT score, 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). As expected, given that our annotators were undergraduate psychology Research Assistants, annotator Moral Foundations and political ideology skew liberal (See Table 2). However, the annotators' MFQ scores nonetheless do exhibit moderate.

[1] This tool is available at https://github.com/limteng-rpi/moral_annotation_tool

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.

Table 2

Annotator Moral Values and Political Ideology

Moral Foundations			Political Ideology	
	Mean	SD		N
Care	3.67	0.70	Very liberal	2
Fairness	3.55	0.66	Liberal	5
Authority	1.96	0.75	Slightly liberal	3
Loyalty	1.86	0.76	Moderate	1
Purity	1.46	0.86	Slightly conservative	1
-	-	-	Conservative	1
-	-	-	Very Conservative	0

Note. Annotator meta-data for the 13 MFTC annotators. Moral Foundations measured on 0-5 scale.

General Sampling Procedure. To assemble the MFTC, we sampled tweets from larger corpora associated with each of the seven discourse domains (See Table 3). While, as noted above, these domains were selected 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 semi-supervised selection as in Garten et al. (2018); 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.

Table 3

Moral Foundations Twitter Corpus 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	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 US Presidential Election	Tweets posted during the 2016 US Presidential Election	Scraped via Twitter API	Followers of @HillaryClinton, @realDonaldTrump, @NYTimes, @washingtonpost, & @WSJ	5,358
Hurricane Sandy	Tweets related to Hurricane Sandy, a hurricane that caused record damage in the U.S.	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 sub-corpus contained in the MFTC. Sub-corpora 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.

This procedure yielded approximately $500 \times 11 = 5500$ 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 5500.

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 3 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 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.

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
Non-moral	1,744	3,848	1,583	2,502	4,452	895	1618
Total	4, 424	5, 593	5, 257	5, 358	4,961	4, 591	4, 891

Note. All tweets were annotated by at least 3 annotators. Majority vote was defined as $\geq$ 50% of annotators.

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 percent of majority vote moral labels assigned by 50% of an even number of annotators (i.e. that were tied).

To evaluate inter-annotator agreement, we calculated both Fleiss' Kappa for multiple annotators (Fleiss, 1971) as well as prevalence and bias adjusted Fleiss' Kappa (PABAK; Sim & Wright, 2005). Fleiss' Kappa represents the degree of observed agreement among annotators beyond what is 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 inter-annotator agreement for each virtue and vice is reasonably high across discourse domains.

Table 6

Inter-annotator Agreement (PABAK and KAPPA) scores for all datasets and foundations

		All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
All Foundations	KAPPA	0.27	0.16	0.37	0.38	0.29	0.19	0.21	0.27
	PABAK	0.29	0.20	0.48	0.41	0.40	0.52	0.23	0.29
Subversion	KAPPA	0.24	0.19	0.05	0.53	0.23	0.08	0.17	0.24
	PABAK	0.67	0.88	0.62	0.89	0.90	0.96	0.47	0.67
Authority	KAPPA	0.29	0.31	-0.01	0.54	0.18	-0.10	0.19	0.29
	PABAK	0.71	0.83	0.85	0.90	0.89	0.57	0.67	0.71
Cheating	KAPPA	0.42	0.25	0.27	0.49	0.41	0.16	0.36	0.42
	PABAK	0.75	0.65	0.70	0.73	0.79	0.88	0.68	0.75
Fairness	KAPPA	0.33	0.31	0.17	0.53	0.44	0.03	0.33	0.33
	PABAK	0.85	0.67	0.85	0.80	0.80	0.94	0.77	0.85
Harm	KAPPA	0.46	0.20	0.18	0.39	0.30	0.37	0.35	0.46
	PABAK	0.65	0.49	0.77	0.61	0.76	0.88	0.77	0.65

Care	KAPPA	0.44	0.25	0.33	0.42	0.32	0.10	0.28	0.44
	PABAK	0.63	0.64	0.88	0.82	0.80	0.97	0.85	0.63
Betrayal	KAPPA	0.18	0.06	0.24	0.36	0.17	0.13	0.18	0.18
	PABAK	0.82	0.88	0.65	0.90	0.90	0.90	0.69	0.82
Loyalty	KAPPA	0.20	0.23	0.32	0.64	0.22	0.11	0.33	0.20
	PABAK	0.62	0.77	0.77	0.87	0.84	0.90	0.80	0.62
Purity	KAPPA	0.16	0.20	0.23	0.28	0.19	0.10	0.28	0.16
	PABAK	0.91	0.91	0.95	0.91	0.76	0.98	0.88	0.91
Degradation	KAPPA	0.19	0.19	0.11	0.27	0.22	0.07	0.28	0.19
	PABAK	0.89	0.87	0.94	0.88	0.90	0.80	0.52	0.89
Non-Moral	KAPPA	0.33	0.02	0.57	0.32	0.29	0.21	0.36	0.33
	PABAK	0.62	0.16	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).

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. Though 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 – i.e. 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.

Methodology

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 et al., 1997; James et al., 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[2] (Graham et al., 2009), a set of a priori selected words associated with each virtue and vice, to

[2] Available at https://www.moralfoundations.org/othermaterials

obtain message-level frequencies for words associated with each virtue and vice. These word-counts were then used to train separate linear Support Vector Machine (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 Moral Foundations Dictionary with the Moral Foundations Dictionary 2 (MFD2)[3] (Frimer, Boghrati, Haidt, Graham, & Dehghani, 2015), 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 Distributed Dictionary Representation (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 multi-task Long Short-Term Memory (LSTM; for an informal introduction to LSTMs, see Olah) 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

[3] Available at https://osf.io/ezn37/

dependencies among words. Here, we use a multitask architecture, which involves training a model to predict labels for multiple outcomes. Specifically, for each discourse domain, we trained a multi-task 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 ≥ 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 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, 8, 9, 10, 11, 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 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.

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	0.61 (0.01)	0.60 (0.04)	0.51 (0.03)	0.67 (0.02)	0.56 (0.03)	0.14 (0.03)	0.60 (0.04)	0.56 (0.03)
	Precision	0.52 (0.01)	0.73 (0.02)	0.61 (0.03)	0.88 (0.03)	0.71 (0.04)	0.93 (0.05)	0.84 (0.03)	0.42 (0.04)
	Recall	0.75 (0.01)	0.51 (0.04)	0.44 (0.04)	0.54 (0.03)	0.46 (0.03)	0.08 (0.02)	0.46 (0.04)	0.85 (0.02)
SVM-MFD2	F1	0.66 (0.01)	0.62 (0.02)	0.57 (0.02)	0.69 (0.02)	0.60 (0.03)	0.13 (0.04)	0.69 (0.02)	0.70 (0.02)
	Precision	0.58 (0.01)	0.54 (0.03)	0.59 (0.03)	0.88 (0.02)	0.74 (0.03)	0.77 (0.23)	0.85 (0.03)	0.59 (0.03)
	Recall	0.75 (0.01)	0.74 (0.02)	0.54 (0.04)	0.57 (0.03)	0.51 (0.04)	0.07 (0.02)	0.57 (0.03)	0.86 (0.02)
SVM-DDR	F1	0.71 (0.01)	0.65 (0.03)	0.62 (0.03)	0.79 (0.01)	0.71 (0.02)	0.14 (0.04)	0.78 (0.01)	0.75 (0.02)
	Precision	0.70 (0.01)	0.72 (0.02)	0.54 (0.03)	0.89 (0.02)	0.71 (0.03)	0.46 (0.10)	0.84 (0.01)	0.71 (0.02)
	Recall	0.73 (0.01)	0.59 (0.03)	0.75 (0.04)	0.72 (0.02)	0.72 (0.03)	0.08 (0.03)	0.73 (0.03)	0.81 (0.02)
LSTM	F1	0.80 (0.01)	0.76 (0.02)	0.69 (0.03)	0.89 (0.01)	0.77 (0.01)	0.14 (0.03)	0.81 (0.02)	0.86 (0.01)
	Precision	0.81 (0.01)	0.77 (0.03)	0.81 (0.03)	0.86 (0.02)	0.78 (0.04)	0.49 (0.14)	0.78 (0.04)	0.97 (0.01)
	Recall	0.79 (0.01)	0.76 (0.02)	0.61 (0.04)	0.92 (0.02)	0.76 (0.04)	0.08 (0.02)	0.84 (0.02)	0.77 (0.02)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

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.

Table 8

Model F1, Precision, and Recall scores for Care

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	0.51 (0.02)	0.54 (0.04)	0.23 (0.04)	0.59 (0.03)	0.52 (0.05)	0.06 (0.02)	0.53 (0.04)	0.54 (0.03)
	Precision	0.49 (0.02)	0.53 (0.05)	0.16 (0.03)	0.65 (0.03)	0.49 (0.05)	0.93 (0.09)	0.50 (0.06)	0.43 (0.04)
	Recall	0.53 (0.03)	0.55 (0.04)	0.40 (0.06)	0.54 (0.04)	0.55 (0.06)	0.03 (0.01)	0.57 (0.06)	0.72 (0.04)
SVM-MFD2	F1	0.56 (0.02)	0.59 (0.03)	0.25 (0.06)	0.64 (0.03)	0.56 (0.05)	0.06 (0.02)	0.53 (0.06)	0.69 (0.03)
	Precision	0.64 (0.02)	0.65 (0.05)	0.17 (0.04)	0.61 (0.04)	0.48 (0.05)	0.89 (0.08)	0.47 (0.07)	0.68 (0.04)
	Recall	0.49 (0.02)	0.55 (0.03)	0.53 (0.10)	0.68 (0.04)	0.67 (0.06)	0.03 (0.01)	0.63 (0.05)	0.70 (0.04)
SVM-DDR	F1	0.48 (0.02)	0.55 (0.03)	0.23 (0.04)	0.61 (0.02)	0.48 (0.04)	0.06 (0.02)	0.43 (0.04)	0.69 (0.03)
	Precision	0.69 (0.02)	0.46 (0.03)	0.13 (0.03)	0.52 (0.03)	0.36 (0.03)	0.48 (0.15)	0.31 (0.03)	0.75 (0.03)
	Recall	0.37 (0.02)	0.68 (0.05)	0.71 (0.07)	0.75 (0.03)	0.74 (0.06)	0.03 (0.01)	0.70 (0.08)	0.65 (0.04)
LSTM	F1	0.63 (0.02)	0.65 (0.05)	0.26 (0.04)	0.77 (0.02)	0.61 (0.06)	0.06 (0.02)	0.36 (0.11)	0.78 (0.03)
	Precision	0.81 (0.03)	0.80 (0.05)	0.76 (0.06)	0.86 (0.02)	0.78 (0.04)	0.64 (0.18)	0.69 (0.07)	0.81 (0.03)
	Recall	0.52 (0.02)	0.55 (0.05)	0.16 (0.03)	0.70 (0.03)	0.50 (0.08)	0.03 (0.01)	0.25 (0.10)	0.75 (0.04)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

To address this issue, we have developed the Moral Foundations Twitter Corpus, 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 Black Lives Matter corpus compared to the All Lives Matter 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 meta-data 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 et al., 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 Moral Foundations Twitter Corpus, our goal is to contribute to this culture of openness and thereby help facilitate both applied and methodological advances in the computational social sciences.

Table 9

Model F1, Precision, and Recall scores for Fairness

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	0.47 (0.02)	0.57 (0.04)	0.30 (0.06)	0.52 (0.05)	0.55 (0.06)	0.03 (0.01)	0.42 (0.04)	0.32 (0.06)
	Precision	0.35 (0.02)	0.72 (0.05)	0.26 (0.05)	0.81 (0.03)	0.81 (0.04)	0.92 (0.11)	0.57 (0.06)	0.56 (0.27)
	Recall	0.72 (0.03)	0.48 (0.05)	0.35 (0.08)	0.38 (0.05)	0.42 (0.06)	0.01 (0.00)	0.33 (0.04)	0.28 (0.13)
SVM-MFD2	F1	0.61 (0.01)	0.59 (0.03)	0.39 (0.05)	0.68 (0.05)	0.70 (0.04)	0.02 (0.02)	0.63 (0.04)	0.59 (0.03)
	Precision	0.59 (0.02)	0.56 (0.05)	0.29 (0.04)	0.80 (0.05)	0.71 (0.04)	0.18 (0.25)	0.71 (0.06)	0.59 (0.04)
	Recall	0.63 (0.02)	0.62 (0.03)	0.60 (0.08)	0.60 (0.05)	0.69 (0.06)	0.01 (0.01)	0.58 (0.05)	0.59 (0.05)
SVM-DDR	F1	0.62 (0.01)	0.70 (0.04)	0.40 (0.03)	0.81 (0.02)	0.69 (0.03)	0.02 (0.01)	0.63 (0.03)	0.54 (0.04)
	Precision	0.79 (0.01)	0.63 (0.06)	0.26 (0.03)	0.78 (0.03)	0.59 (0.04)	0.42 (0.25)	0.56 (0.03)	0.85 (0.04)
	Recall	0.51 (0.02)	0.79 (0.03)	0.78 (0.06)	0.85 (0.03)	0.84 (0.04)	0.01 (0.01)	0.72 (0.06)	0.40 (0.04)
LSTM	F1	0.70 (0.01)	0.75 (0.04)	0.43 (0.04)	0.88 (0.02)	0.75 (0.03)	0.02 (0.02)	0.55 (0.07)	0.10 (0.06)
	Precision	0.81 (0.02)	0.84 (0.04)	0.81 (0.07)	0.91 (0.02)	0.85 (0.03)	0.35 (0.22)	0.76 (0.04)	0.06 (0.04)
	Recall	0.61 (0.02)	0.68 (0.04)	0.30 (0.04)	0.86 (0.03)	0.68 (0.06)	0.01 (0.01)	0.43 (0.09)	0.87 (0.19)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

Table 10

Model F1, Precision, and Recall scores for Loyalty

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	0.40 (0.02)	0.32 (0.07)	0.41 (0.04)	0.61 (0.05)	0.33 (0.07)	0.05 (0.07)	0.54 (0.06)	0.35 (0.03)
	Precision	0.38 (0.01)	0.26 (0.07)	0.38 (0.04)	0.69 (0.05)	0.24 (0.05)	0.08 (0.07)	0.58 (0.06)	0.47 (0.04)
	Recall	0.42 (0.03)	0.45 (0.12)	0.44 (0.05)	0.55 (0.06)	0.54 (0.11)	0.05 (0.06)	0.52 (0.06)	0.28 (0.03)
SVM-MFD2	F1	0.41 (0.02)	0.40 (0.06)	0.43 (0.04)	0.68 (0.05)	0.33 (0.06)	0.05 (0.03)	0.53 (0.06)	0.34 (0.03)
	Precision	0.40 (0.02)	0.56 (0.10)	0.38 (0.04)	0.80 (0.05)	0.23 (0.04)	0.22 (0.15)	0.52 (0.07)	0.51 (0.05)
	Recall	0.42 (0.02)	0.32 (0.05)	0.51 (0.06)	0.60 (0.05)	0.58 (0.10)	0.03 (0.02)	0.55 (0.07)	0.26 (0.03)
SVM-DDR	F1	0.36 (0.02)	0.37 (0.03)	0.46 (0.05)	0.73 (0.04)	0.27 (0.04)	0.05 (0.03)	0.52 (0.03)	0.36 (0.04)
	Precision	0.66 (0.01)	0.25 (0.03)	0.34 (0.04)	0.62 (0.06)	0.17 (0.03)	0.53 (0.23)	0.41 (0.04)	0.73 (0.06)
	Recall	0.25 (0.02)	0.74 (0.06)	0.72 (0.08)	0.88 (0.03)	0.75 (0.08)	0.02 (0.01)	0.70 (0.05)	0.24 (0.03)
LSTM	F1	0.43 (0.02)	0.38 (0.09)	0.50 (0.03)	0.87 (0.03)	0.26 (0.03)	0.03 (0.01)	0.38 (0.07)	0.40 (0.05)
	Precision	0.77 (0.03)	0.69 (0.09)	0.77 (0.04)	0.92 (0.03)	0.71 (0.05)	0.75 (0.18)	0.73 (0.07)	0.71 (0.07)
	Recall	0.30 (0.03)	0.26 (0.08)	0.37 (0.03)	0.83 (0.07)	0.16 (0.02)	0.02 (0.01)	0.26 (0.07)	0.28 (0.05)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

Table 11

Model F1, Precision, and Recall scores for Authority

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	0.42 (0.01)	0.55 (0.07)	0.16 (0.04)	0.73 (0.04)	0.42 (0.07)	0.00 (0.00)	0.39 (0.05)	0.43 (0.05)
	Precision	0.48 (0.02)	0.43 (0.07)	0.10 (0.03)	0.63 (0.05)	0.31 (0.06)	0.13 (0.27)	0.39 (0.06)	0.41 (0.05)
	Recall	0.38 (0.02)	0.77 (0.10)	0.36 (0.07)	0.86 (0.03)	0.70 (0.05)	0.00 (0.00)	0.39 (0.05)	0.45 (0.06)
SVM-MFD2	F1	0.40 (0.02)	0.41 (0.07)	0.19 (0.04)	0.68 (0.06)	0.38 (0.04)	0.01 (0.01)	0.38 (0.04)	0.40 (0.03)
	Precision	0.53 (0.03)	0.74 (0.08)	0.12 (0.03)	0.57 (0.07)	0.26 (0.04)	0.58 (0.38)	0.39 (0.04)	0.69 (0.04)
	Recall	0.33 (0.02)	0.29 (0.06)	0.57 (0.12)	0.85 (0.03)	0.68 (0.04)	0.00 (0.00)	0.38 (0.04)	0.29 (0.03)
SVM-DDR	F1	0.36 (0.02)	0.48 (0.05)	0.19 (0.03)	0.73 (0.04)	0.30 (0.04)	0.01 (0.01)	0.47 (0.03)	0.50 (0.04)
	Precision	0.71 (0.02)	0.33 (0.04)	0.11 (0.02)	0.60 (0.06)	0.18 (0.03)	0.50 (0.25)	0.37 (0.03)	0.78 (0.04)
	Recall	0.24 (0.02)	0.85 (0.06)	0.73 (0.08)	0.93 (0.04)	0.81 (0.06)	0.01 (0.00)	0.65 (0.05)	0.37 (0.04)
LSTM	F1	0.47 (0.02)	0.57 (0.07)	0.19 (0.02)	0.83 (0.03)	0.33 (0.07)	0.01 (0.01)	0.47 (0.03)	0.59 (0.03)
	Precision	0.80 (0.02)	0.85 (0.07)	0.77 (0.07)	0.91 (0.05)	0.80 (0.09)	0.24 (0.31)	0.67 (0.06)	0.80 (0.06)
	Recall	0.34 (0.02)	0.43 (0.07)	0.11 (0.01)	0.76 (0.06)	0.21 (0.06)	0.01 (0.01)	0.36 (0.03)	0.46 (0.03)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

Table 12

Model F1, Precision, and Recall scores for Purity

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
SVM-MFD	F1	0.30 (0.03)	0.15 (0.01)	0.07 (0.02)	0.54 (0.06)	0.35 (0.05)	0.03 (0.01)	0.45 (0.06)	0.20 (0.10)
	Precision	0.43 (0.04)	0.08 (0.01)	0.04 (0.01)	0.47 (0.08)	0.29 (0.06)	0.96 (0.06)	0.47 (0.06)	0.38 (0.17)
	Recall	0.23 (0.03)	0.82 (0.06)	0.43 (0.13)	0.64 (0.09)	0.45 (0.05)	0.02 (0.00)	0.44 (0.06)	0.14 (0.08)
SVM-MFD2	F1	0.33 (0.02)	0.34 (0.03)	0.13 (0.06)	0.49 (0.07)	0.43 (0.07)	0.02 (0.01)	0.51 (0.04)	0.20 (0.06)
	Precision	0.59 (0.03)	0.73 (0.07)	0.07 (0.04)	0.36 (0.06)	0.33 (0.06)	0.30 (0.35)	0.50 (0.04)	0.64 (0.16)
	Recall	0.23 (0.02)	0.22 (0.02)	0.54 (0.13)	0.76 (0.10)	0.64 (0.07)	0.01 (0.01)	0.53 (0.05)	0.12 (0.03)
SVM-DDR	F1	0.24 (0.02)	0.25 (0.04)	0.11 (0.03)	0.34 (0.07)	0.33 (0.05)	0.03 (0.01)	0.54 (0.05)	0.14 (0.03)
	Precision	0.66 (0.03)	0.15 (0.03)	0.06 (0.02)	0.21 (0.05)	0.22 (0.04)	0.35 (0.16)	0.45 (0.04)	0.74 (0.13)
	Recall	0.15 (0.02)	0.76 (0.12)	0.88 (0.09)	0.84 (0.09)	0.72 (0.07)	0.01 (0.01)	0.69 (0.06)	0.08 (0.02)
LSTM	F1	0.41 (0.02)	0.57 (0.07)	0.07 (0.03)	0.48 (0.10)	0.47 (0.05)	0.04 (0.02)	0.53 (0.07)	0.15 (0.03)
	Precision	0.80 (0.03)	0.85 (0.07)	0.81 (0.24)	0.81 (0.10)	0.79 (0.08)	0.48 (0.19)	0.71 (0.08)	0.72 (0.10)
	Recall	0.28 (0.02)	0.43 (0.07)	0.03 (0.02)	0.34 (0.10)	0.33 (0.04)	0.02 (0.01)	0.43 (0.07)	0.09 (0.02)

Notes. All models were fit with 10-fold cross-validation. Metrics indicate mean performance across folds. Parenthetical numbers indicate SDs across folds.

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*.

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 (28), 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).

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*.

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*.

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), 1–14.

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).

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 (44), 11203–11208.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5), 378.

Frimer, J., Boghrati, R., Haidt, J., Graham, J., & Dehghani, M. (2015). Moral foundations dictionary 2.0.

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 Proc. IJCAI 2016 workshop on computational modeling of attitudes*.

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 (1), 344–361.

Garten, J., Kennedy, B., Hoover, J., Sagae, K., & Dehghani, M. (2019). Incorporating demographic embeddings into language understanding. *Cognitive science*, 43 (1), e12701.

Garten, J., Kennedy, B., Sagae, K., & Dehghani, M. (2019). Measuring the importance of context when modeling language comprehension. *Behavior research methods*, 1–13.

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). 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 (5), 1029–1046.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9 (8), 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*. 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 (1), 9.

Hoover, J., Johnson-Grey, K., Dehghani, M., & Graham, J. (2017). Moral values coding guide.

Iliev, R., Dehghani, M., & Sagi, E. (2014, July). Automated text analysis in psychology: methods, applications, and future developments. *Language and cognition*, 1–26.

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

Johnson, K., & Goldwasser, D. (2018). Classification of moral foundations in microblog political discourse. *In Proceedings of the 56th annual meeting of the association for computational linguistics* (volume 1: Long papers) (Vol. 1, pp. 720–730).

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 (4), 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 The 2018 IEEE/ACM international conference on advances in social networks analysis and mining* (ASONAM2018).

Lovett, B. J., Jordan, A. H., & Wiltermuth, S. S. (2012, July). Individual differences in the moralization of everyday life. *Ethics & behavior*, 22 (4), 248–257.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. *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 (6), 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*.

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.

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).

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. American anthropologist, 88(2), 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 (2), 132–144.

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

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 (2), 147–168.

Weller, S. C., & Mann, N. C. (1997). Assessing rater performance without a" gold standard" using consensus theory. Medical Decision Making, 17(1), 71-79.

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.