Exploring Morality in Argumentation

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, Heiner Stuckenschmidt
University of Mannheim
{jonathan,ines,ioana,heiner}@informatik.uni-mannheim.de

Abstract

Sentiment and stance are two important concepts for the analysis of arguments. We propose to add another perspective to the analysis, namely *moral sentiment*. We argue that moral values are crucial for ideological debates and can thus add useful information for argument mining. In the paper, we present different models for automatically predicting moral sentiment in debates and evaluate them on a manually annotated test set. We then apply our models to investigate how moral values in arguments relate to argument quality, stance, and audience reactions.

1 Introduction

Argumentation mining is a new research field that is closely related to the subfields of stance detection and sentiment analysis (Stede, 2020), since "every argument carries a stance towards its topic, often expressed with sentiment".[1] In addition to stance and sentiment, there is another dimension that can play an important role in debates, namely *moral beliefs*. A debater's moral beliefs go beyond stance and can be expressed with varying sentiment. They play an important role in ideological debates and cannot be resolved by simply comparing facts but often involve a battle of ideas and a clash of different belief systems. Consider the following arguments on whether or not gay marriage should be legal.[2]

(1) The institution of marriage has traditionally been defined as being between a man and a woman.

(2) Denying some people the option to marry is discriminatory and creates a second class of citizens.

Both arguments are based on moral belief systems. The first argument refers to moral values that promote respect for tradition, while the second focuses on fairness and equal rights. Arguments that express an opposite stance towards the topic usually differ concerning their moral framing. On the other hand, we observe that arguments expressing a similar stance towards a certain topic may still differ with regard to how the argument is framed, as illustrated in example (3) and (4) below. While (3) opposes the legalization of prostitution because it is considered as a harmful form of oppression targeting women, example (4) depicts prostitution as increasing the danger of diseases and contamination. This makes moral framing an interesting ingredient for argument mining.

(3) Prostitution and human trafficking are forms of gender-based violence.

(4) Prostitution is the biggest vector of sexually transmitted diseases.

In the paper, we argue that identifying moral values in debates has the potential to support argument analysis and to help with different subtasks related to argument mining. Being able to distinguish between arguments with similar stance and sentiment but framed according to different moral categories can help to identify new arguments and can improve camp detection, thus supporting more fine-grained modeling of debaters beyond stance. Furthermore, moral framing is of particular interest for the analysis of political debates (Lakoff, 1997; Roggeband and Vliegenthart, 2007).

[1] Cited from the workshop website (https://argmining2020.i3s.unice.fr/).

[2] From https://gaymarriage.procon.org (accessed August 25, 2020).

Proceedings of the 7th Workshop on Argument Mining, pages 30–40
Barcelona, Spain (Online), December 13, 2020.

In practice, however, predicting moral sentiment from text poses several challenges. First, morality is a fuzzy concept, and it is difficult to find an operationalization that turns it into measurable data. Moral sentiment is often expressed implicitly and thus hard to detect, based merely on the presence of lexical cues. In addition, human coders might be biased by their own belief systems, which casts doubt on the validity of the annotations used to train or evaluate automatic systems.

In the paper, we present work in progress where we evaluate different models for the prediction of moral framing in text on two datasets and assess the benefits of these predictions for the analysis of arguments. Based on three datasets with argumentative text, we investigate whether we can find correlations between moral values and different aspects of argumentation, such as argument quality, stance, or audience approval. We are interested in the following research questions:

RQ1: Do debaters that produce high-quality arguments make more or less frequent use of moral framing?
RQ2: Is moral framing more strongly related to a positive or negative stance?
RQ3: Can we find a positive correlation between the audiences' approval and the use of moral frames?

Our main contributions are the following: (i) We augment the ArgQuality Corpus of Wachsmuth et al. (2017) with annotations for moral values, as a first test set for the evaluation of moral sentiment in argumentation; (ii) We evaluate two methods for the prediction of moral sentiment on the new dataset; (iii) We present a correlation study investigating the relation between moral framing and argument quality, stance, and audience reactions.

The paper is structured as follows. We first review work on quantifying moral sentiment in text (§2). In §3, we describe the annotation of our test set and present different approaches to the automatic detection of moral sentiment in debates. Then we discuss our correlation analysis (§4), and in §6 we summarise our results and conclude.

2 Related Work

As an operationalisation for the concept of moral sentiment, we refer to Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Graham et al., 2013). MFT has its roots in social and cultural psychology and assumes the existence of innate and universally available psychological systems that build the foundations of intuitive ethics. These foundations are augmented by culture-specific constructs of virtues and backed up by personal narratives "that people construct to make sense of their values and beliefs" (Graham et al., 2013)[p.17], and are also reinforced by institutional environments. MFT assumes that all moral issues can be described along the following dimensions: *Care-Harm, Fairness-Cheating, Loyalty-Betrayal, Authority-Subversion*, and *Purity-Degradation*.[3]

Dictionary-based approaches to MFT The first version of the Moral Foundations Dictionary (MFD) was presented by Graham et al. (2009) and has been used for a content analysis of christian sermons held in liberal and conservative congregations. Each of the five moral foundations listed above has been further split into a *vice* and a *virtue* subcategory, reflecting the positive and negative ends of each dimension. Examples are *peace*, protect*, compassion** for $Care_{virtue}$ and *suffer*, crush*, killer** for $Care_{vice}$. The MFD includes, on average, 32 words per moral subcategory. Frimer et al. (2019) presents a new version of the MFD with more entries per MF subcategory, selected according to prototypicality estimates for each MF, based on cosine similarity for Word2Vec embeddings for each item in the dictionary. While the authors admit that the construct validity of the MFD 2.0 is not better than for the original MFD, they recommend the use of the MFD 2.0 due to its improved coverage. Other work on expanding the MFD includes Rezapour et al. (2019) who increase the original size of the dictionary to over 4,600 lexical items, using a quality controlled, human in the loop process.[4]

The MF dictionary has been used in several studies in the political and social sciences, psychology, and related fields (Takikawa and Sakamoto, 2017; Matsuo et al., 2018; Lewis, 2019). Dictionary-based approaches to measuring moral values in text, however, have severe shortcomings. They can neither

[3] Another foundation currently under discussion is *Liberty-Opression*.

[4] Arguably, a comparison of size is not meaningful, given that the original MFD includes regular expressions that can match an unknown number of instance types while the expanded lexicon includes word forms for unigrams and word compounds.

	Care	*Fairness*	*Loyalty*	*Authority*	*Purity*	*Moral*
Cohen's κ	.469	.407	.529	.363	.280	.434
Krippendorff's α	.459	.400	.530	.356	.255	.402
Absolute Agreement (positive/negative)	60/187	16/267	10/294	12/274	13/257	165/68
Absolute Disagreement	73	37	16	34	50	87

Table 1: Inter-Annotator Agreement for the five MFs and for a binary label (*Moral*: yes/no).

account for unknown words or the different meanings a word can take, nor do they consider that shifter words and negation can change the polarity of an expression. In addition, we expect that moral vocabulary might vary considerably, depending on the speaker's age and other geopolitical, social, and cultural variables. Garten et al. (2016) address the coverage problem of dictionary-based approaches by replacing the terms in the MF dictionary with their averaged vector representations in distributional space. They show that predicting moral foundations based on the cosine similarity of the words in a text to the distributional representations outperforms a naive method that predicts MF based on word counts.

Machine learning-based approaches Recent work has applied the framework of MFT to research questions in the social and political sciences (Fulgoni et al., 2016; Johnson and Goldwasser, 2018; Rezapour et al., 2019; Xie et al., 2019), replacing dictionary-based counts with more sophisticated methods. Johnson and Goldwasser (2018) model moral framing in politicians' tweets, using probabilistic soft logic (Bach et al., 2013). Lin et al. (2018) improve the prediction of moral foundations by acquiring additional background knowledge from Wikipedia, using information extraction techniques such as entity linking and cross-document knowledge propagation. Xie et al. (2019) study the change of moral sentiment in longitudinal data, presenting a parameter-free model that predicts moral sentiment on three different levels: (i) moral relevance, (ii) moral polarity, and (iii) the ten moral subclasses of the MFD encoding the virtue/vice dimension for each MF. Finally, Resapour et al. (2019) show that using dictionary counts for moral sentiment as features in a supervised classification setup can increase results for stance detection.

3 Predicting Moral Sentiment in Tweets and Debates

3.1 A New Test Set for Moral Framing in Argumentation

As a test set for evaluating moral framing in English argumentative text, we use the Dagstuhl ArgQuality Corpus (Wachsmuth et al., 2017). The dataset contains 320 arguments with approx. 22,600 tokens, covering 16 topics, and is balanced for stance. The data was extracted from two online debate platforms by Habernal and Gurevych (2016).[5] Each instance has been annotated by three coders, using a fine-grained scheme to assess the arguments' quality. The data also provides a majority score for each dimension of argument quality (Wachsmuth et al., 2017). The authors report a low agreement for the individual annotations (.51 Krippendorff's α) but a high majority agreement (94%).

We further augment the ArgQuality corpus with annotations for moral foundations, manually coded by two of the authors.[6] We chose not to annotate the 10 subclasses encoded in the dictionary but considered the two ends of each dimension (virtue/vice) as one category. The motivation behind this decision is that both are closely related, and it is often unclear which end of the dimension is addressed, particularly for negated sentences. E.g., "I could never hurt you" could either be considered as an instance of Harm as it uses vocabulary related to this dimension or could be annotated as the opposite, Care, as it talks about *not* being able to harm somebody, thus being more strongly related to the *virtue* class.

Table 1 shows inter-annotator agreement (IAA) scores for individual MFs on the ArgQuality dataset. As expected, IAA is low, being roughly in the same range as agreement scores reported for the annotation of emotions (Schuff et al., 2017; Wood et al., 2018), thus giving evidence for the subjectivity of the task. Our IAA is not directly comparable to Hoover et al. (2020) as they report Fleiss' κ for the 10 fine-grained subclasses, with an avg. of .315 κ over all 10 classes.

[5] `www.createdebate.com` and `convinceme.net`.

[6] All resources created for this paper are available at `https://github.com/dwslab/Morality-in-Arguments`

3.2 Methods for the Prediction of Moral Sentiment

We model moral sentiment prediction as a text classification task and propose two distinct methods for feature generation. The first method is based on a *sense-disambiguated* version of the MFD and extends its coverage by exploiting relations in Wordnet (WN) (Fellbaum, 2010). The second method uses the MFD as seed data to learn BERT sequence embeddings that encode moral sentiment. The representations created by each method are fixed-sized vectors that can easily be combined by concatenation.

I. Sense-disambiguated features (WN-PPR) The MFD has two main disadvantages that we try to overcome with this method. First, the lexicon contains many words with different word senses, where the moral value only applies to one specific sense. Thus, we link the dictionary entries to their corresponding WN synsets. This way, *fair* is only considered to be related to the MF *fairness-cheating* if used as synonym to *just* or *honest*, but not if used as synonym to *carnival, funfair* or *attractively feminine*. Also, this way we overcome the problems resulting from the use of regular expressions in the MFD (e.g., *defenestration* would belong to the MF *Care* because of the entry *defen**, and *Churchill* would trigger *Purity* because of the entry *church**). The second disadvantage of the MFD is its low coverage, which we extend by running Personalized Pagerank (Haveliwala, 2003) on the set of WN synsets that have been linked to dictionary entries.

a) Linking MFD entries to WN To create a word sense disambiguated version of the MFD, one expert annotator was presented with the following information: a specific moral foundation; a WN synset whereof at least one word in the synset is part of the respective MF in the MFD; and its definition. With this information, the annotator decided whether the synset is relevant for the moral foundation in question or not. Overall, the resulting lexicon contains, on average, 61 synsets per MF.

b) Extending the disambiguated Lexicon We extend the disambiguated lexicon by exploiting relations between synsets in WN, such as *hypernym* or *similar to*. Concretely, we run personalized pagerank on the graph consisting of the WN synsets and the relations between them for every MF, using the corresponding lexicon entries as seed nodes. This way, each WN synset is assigned a fixed-sized vector containing scores for each MF, including the category *GeneralMorality*.[7] We expect that higher scores reflect a stronger correspondence between the synset and the respective MF.

c) Extracting features from text Given a short English text, we first extract WN 3.0 synsets using the disambiguation method by Tan (2014). Then we link these synsets to WN 3.1, using the official Wordnet Search Engine[8] and, if necessary, resolving the final mapping manually. For instance, a variety of offensive terms have been removed in WN 3.1, and thus, we had to link terms like *darky* or *tom* to *black (noun.person)* manually.[9] As each of the synsets is assigned a fixed-size score vector in our lexicon, any function to aggregate these vectors is conceivable. To obtain vectors that do not depend on the input text's length, we decided to take their mean. The result is a vector consisting of 6 entries, where each entry represents a MF, including *GeneralMorality*.

II. Contextualized MF sequence embeddings (SBERT-Wiki) Our second method uses Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to obtain text representations that encode moral sentiment. SBERT is a modification of the original transformer model, based on siamese networks. The modified SBERT encodes sentence similarity in a human interpretable way where similar sentence pairs can be retrieved efficiently, based on cosine similarity. While previous work has computed BERT-based embeddings for text sequences (i) by averaging (or summing) over all word embeddings for this particular text or (ii) by using the network output at the position of the [CLS] token, Reimers and Gurevych (2019) show that those representations are not well suited to encode sentence similarity and often yield inferior results as compared to using averaged Glove embeddings (Pennington et al., 2014).

[7] *GeneralMorality* includes terms related to moral concepts that do not fit into one of the five MFs (e.g. *ethic, good, evil*).

[8] `http://wordnet-rdf.princeton.edu/json/pwn30/...`

[9] We are aware that this treatment is not optimal. A better solution would link those terms to a synset that captures their offensive usage, similar to the one for *Kraut: offensive term for a person of German descent*.

SBERT can also be applied to tasks where an anchor text is compared to a positive and a negative text sample, thus learning to maximize a score based on the similarity of the anchor text to the first sentence (the positive sample) and its distance to the second text (the negative sample). For that, SBERT uses the triplet objective function (equation 1) where d is a distance metric (here: Euclidean distance), and the margin is set to 1.

$$L = max(d(s_{anchor} - s_{pos}) - d(s_{anchor} - s_{neg}) + margin, 0) \quad (1)$$

We fine-tune SBERT embeddings so that they encode different moral foundations. First, we download all short Wikipedia abstracts from DBpedia[10] and label them with their corresponding MF (if any), using weak supervision. Our approach is based on the MFD and proceeds as follows: For each dictionary entry, we search in Wikipedia for corresponding articles to get a set of candidates consisting of articles whose title is a lexicon entry (including redirections) and articles that are linked by the lexical entry's disambiguation page. From these candidates, we manually select the ones related to the MF and label their abstracts accordingly.

This approach yields 317 short abstracts from Wikipedia, labeled with moral foundations, extracted from a pool of 4,935,596 unlabelled short Wikipedia abstracts. We iterate over each abstract in the annotated dataset, considering the abstract as the anchor text. First, we retrieve all other abstracts labeled with the same moral foundation as the anchor and create pairs of (anchor, positive sample). Then, for each pair, we randomly select 3 labeled abstracts that belong to a different moral foundation as well as 7 abstracts from the unlabelled pool as negative samples, assuming that the unlabelled abstracts also do, more often than not, either belong to a different moral foundation or do not express any moral content. This gives us a total of 10 negative samples for each pair and results in a weakly supervised dataset with 107,940 instances. We then fine-tune the model on the data, using the same settings as reported in Reimers and Gurevych (2019). After the training is completed, we use the learned model to retrieve representations for new text sequences from different argumentation datasets, expecting that the fine-tuned embeddings will now capture some aspects of moral sentiment. We compare our approach with the pretrained SBERT embeddings (bert-base-nli-stsb-mean-tokens) of Reimers and Gurevych (2019), trained without the fine-tuning step on the weakly supervised Wikipedia abstracts.

3.3 State of the Art and Baselines

multi-label BERT To compare our lexicon-based methods with a state-of-the-art approach to text classification, we train a multi-label text classifier based on BERT. We use a publicly available implementation in pytorch[11] that replaces the cross-entropy loss with a binary cross-entropy with logits to adapt the BERT sequence classifier to the multi-label setup.

The model includes an input embedding layer for the pretrained BERT embeddings, the BERT encoder with 12 attention layers, and, as final layer, a linear transformation, with one dimension for each class. This gives us six output dimensions: the five moral foundations + one class for tweets with non-moral content. Our model uses the pretrained English uncased BERT base embeddings with a vocabulary size of around 30,000. We use the same data splits and preprocessing in all experiments (for details, see §3.4). In contrast to our other models, however, BERT further segments the input text into subword tokens (WordPiece tokenization), which might increase coverage for words not seen in the training data.

Random baseline The *Random* baseline assigns labels randomly but respecting the class distribution in the training data. Results are averaged over 100 trials.

MFD baseline Given a text, we compute frequency counts for each MF, based on the entries in the MFD, and normalize by text length. We use these count-based vectors as features for the text classifier. Similar to WN-PPR, we derive one feature per MF, including general morality.

[10] `https://wiki.dbpedia.org/downloads-2016-10`

[11] The code was adapted from `https://medium.com/huggingface/multi-label-text-classification-using-bert-the-mighty-transformer-69714fa3fb3d` and is based on the HuggingFace library (`https://github.com/huggingface/pytorch-pretrained-BERT`).

Method	Moral	Care	Fairness	Loyalty	Authority	Purity	Average (excl. Moral)
Random baseline	.519	.173	.169	.100	.099	.055	.119
MFD baseline	.630	.332	.213	.166	.231	.141	.217
multi-label BERT	.669	**.510**	**.573**	**.437**	**.377**	**.363**	**.452**
WN-PPR	.628	.334	.379	.311	.210	.088	.264
SBERT-Base	.685	.434	.511	.372	.327	.214	.372
SBERT-Wiki	**.697**	.463	.516	.377	.341	.220	.383
WN-PPR + SBERT-Wiki	.689	.446	.520	.387	.346	.230	.386

Table 2: Binary F1-scores on the MFTC for individual MFs (F1 for the positive class). The last column shows the average over the F1 scores for the five MFs (excluding *Moral*).

Method	Moral	Care	Fairness	Loyalty	Authority	Purity	Average (excl. Moral)
Random baseline	.658	.257	.179	.096	.134	.105	.154
MFD baseline	**.853**	.056	.237	.043	.200	.086	.124
multi-label BERT	.444	**.517**	**.519**	**.138**	.157	.208	**.308**
WN-PPR	.756	.118	.253	.049	.105	.029	.111
SBERT-Base	.703	.280	.342	.065	.148	.133	.194
SBERT-Wiki	.730	.339	.246	.125	**.233**	**.318**	.252
WN-PPR + SBERT-Wiki	.686	.298	.351	.067	.040	.135	.178

Table 3: Binary F1-scores on the Dagstuhl ArgQuality Corpus for individual MFs (F1 for the pos. class).

3.4 Data

We now present the data used for the evaluation of the methods described above (§3.2) for the prediction of moral sentiment in tweets and debates. As training data for our MF classifiers, we use the Moral Foundations Twitter Corpus (Hoover et al., 2020), a collection of approximately 35,000 tweets covering seven controversial topical threads: All Lives Matter, Black Lives Matter, the Baltimore protests, the 2016 Presidential election, hate speech & offensive language (Davidson et al., 2017), Hurricane Sandy, and #MeToo. Each tweet has been annotated with MF by at least three trained annotators. The authors report Fleiss' κ and PABAK, a measure adjusted for prevalence and bias (Sim and Wright, 2005). IAA is relatively low (with a Fleiss κ in the range of 0.24 - 0.46 and PABAK ranging from 0.65 - 0.85) and shows considerable variation across the different moral domains and threads.

We follow the procedure described in Hoover et al. (2020) to create a gold standard from the annotated tweets and consider a label as *gold* if it was assigned by at least half of the annotators. Thus, our gold standard includes 6 labels: one for each MF and a sixth one for GeneralMorality. Note that in the MFTC, this label is called *Non-moral* while we report results for its inverse, which we call *Moral*. We normalized the tweets using the script available from the Glove website.[12] We noticed that the dataset includes many near-duplicates (e.g., 96 instances of *homosexuality is a sin*). To ensure that these near-duplicates do not appear in both training *and* test set, we split the data into the different threads and present results for a seven-fold cross-validation where we train the models on six threads and evaluate on the remaining one. We also evaluate the models trained on the MFTC on out-of-domain data from the ArgQuality Corpus, where we consider all labels assigned by each of the two annotators as ground truth.[13]

3.5 Results for MF Prediction on Tweets and Debates

We conduct experiments on the Twitter corpus, testing different traditional classification methods, and report results for the best performing classifiers only. For WN-PPR and MFD-Features, this was a k-nearest-neighbors classifier, and for SBERT-Base, SBERT-Wiki, as well as for WN-PPR + SBERT-Wiki a Linear Discriminant Analysis.[14] All other results, as well as the correlations reported in §4, refer to these classification methods.

Table 2 shows results on the MFTC for our different methods. Not surprisingly, multi-label BERT outperforms all other methods on the Twitter data. However, our lexicon-based methods outperform the random baseline for each category, with the best results obtained by the concatenation of SBERT-Wiki

[12] `https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb`

[13] As the data has been annotated by two of the authors, we can be sure that we do not have to eliminate spammers.

[14] We use the scikit-learn implementation for these methods. Other methods we tried include Logistic Regression, Decision Trees, Naïve Bayes, Support Vector Machines.

with WN-PPR. WN-PPR on its own only yields poor results, barely outperforming the constant and the MFD baseline.

When applying the classifiers to the out-of-domain data from the ArgQuality corpus (Table 3), multilabel-BERT still yields best results, but now SBERT-Wiki outperforms BERT on the *Authority, Purity* and *Moral* categories. The lower performance for the *Moral* class can be explained by the differences in class distribution between the two datasets. In the MFTC, this class makes up for approximately 57% of the training instances, while the amount of moral instances in the debate corpus is much higher (79%). The lexicon-based methods are not sensitive to the class distribution in the training data, which, in this case, makes them more robust. Still, all systems fail to beat the majority baseline for the Moral class which has a binary F1-score of 0.881.[15] WN-PPR again performs poorly with results below the Random baseline, and results for the MFD baseline also fail to outperform Random. This time, results for the concatenation of WN-PPR and BERT-Wiki are considerably worse than for BERT-Wiki alone.

4 Correlation Studies

To study the impact of moral framing in argumentation, we investigate the correlation between moral sentiment and other properties of argumentative text, namely argument quality, stance, and audience reactions. For this, we use the multilabel-BERT model that yielded the best results on both datasets.

4.1 Data

The **Dagstuhl ArgQuality Corpus** contains arguments that are annotated with different dimensions of argument quality, such as *cogency* and *credibility*, as well as a score for *overall quality*. Some of the dimensions are also interesting for contexts other than argument quality, such as *clarity* and *emotional appeal*.

The **IBM Argument Quality Ranking Corpus** (Gretz et al., 2019) is used to triangulate our findings on the Dagstuhl ArgQuality Corpus and to investigate the correlation between moral sentiment and an argument's stance. The corpus contains more than 30,000 arguments on 71 topics, labelled for quality (*good* or *bad*) and stance (*pro* or *con*) by crowd annotators. To obtain ranks for argument quality, the authors apply two different strategies, which both give more weight to the answers of reliable annotators.

We use **CORPS** (Guerini et al., 2013) to investigate whether moral sentiment in political speeches has an impact on the audience. CORPS includes >3,600 political speeches held by more than 203 different speakers, tagged for audience reactions such as *applause, laughter* or *booing*. The motivation for creating the corpus was that such tags might highlight passages in the speech where an attempt has been made by the speaker to persuade the audience, either successful or not. We expect to find a correlation between text passages that triggered a positive audience reaction (i.e. *applause*) and moral framing, but not for *laughter* (we focus on these two tags as the other tags are relatively rare in comparison[16]). We also exclude mixed tags that mark two different reactions for the same text passage (*laughter; applause*). To test our hypothesis, we predict moral sentiment for the speech passages directly before an audience reaction was triggered. We consider up to 360 tokens of speech context and omit all speech passages where another tag occurs within this context.

4.2 Results for the Correlation Analysis

Table 4 shows Spearman correlations between argument quality, stance, and audience reactions and a) human annotations (HU) and b) labels predicted by multi-label BERT (BM). We observe a weak positive correlation between argument quality and moral sentiment for the two most frequent categories (*Moral, Care*) on the ArgQuality data. For the other MFs, there are no significant effects. On the IBM-AQR Corpus, we see a consistent and significant positive correlation for *Care* and *Fairness*. However, the effect is very weak. For the subdimensions of argument quality, the correlations tend to be similar to the ones for overall quality and are highest for *emotional appeal*, which seems plausible. Concerning argument stance, we again find slightly positive correlations for *Care* and *Moral*. Results on the CORPS

[15] The majority baseline is not included in tables 2 and 3 because its binary F1-score is zero for all classes except *Moral*.

[16] Applause: 23,095; Laughter: 5,857; Booing: 532; Cheers: 80; Sustained applause: 61; Spontaneous demonstration: 16.

	Care		Fairness		Loyalty		Authority		Purity		Moral	
	HU	BM	HU	BM	HU	BM	HU	BM	HU	BM	HU	BM
Dagstuhl ArgQuality Corpus												
overall quality	**.25**	**.15**	.10	.08	.05	**.10**$^{-}$	-.09	.05^{+}	.03	.07^{-}	**.19**	**.21**
local acceptability	**.18**	.09	.00	-.04^{+}	.00	.04^{-}	**-.15**	-.01^{+}	-.03	.07^{-}	.03	.09
appropriateness	**.30**	**.17**	-.01	.03	-.02	.05^{-}	-.09	.00^{+}	.01	.02	**.19**	**.15**
arrangement	**.24**	**.16**	.08	.03	.03	.08^{-}	-.06	-.01^{+}	.04	.05	**.16**	**.17**
clarity	**.17**	**.17**	.02	.02	.05	**.12**$^{-}$	-.03	-.01^{+}	.03	.06	.09	**.21**$^{-}$
cogency	**.24**	**.16**	.05	.06	-.02	.03^{-}	-.10	.05^{+}	.01	.03^{-}	**.10**	**.18**
effectiveness	**.25**	**.17**	.09	.09	-.05	.07^{-}	-.10	-.02^{+}	.04	.04	**.21**	**.17**
global acceptability	**.23**	**.12**	.05	.05	-.01	.07^{-}	**-.12**	.02^{+}	.01	.04	**.12**	**.13**
global relevance	**.15**	.06	**.11**	.09	.02	.07^{-}	-.11	.00^{+}	.04	.05	**.12**	**.11**
global sufficiency	**.19**	.11	**.11**	.11	-.01	.06^{-}	-.04	-.03^{+}	.07	.05^{-}	**.19**	**.14**$^{+}$
reasonableness	**.23**	**.17**	.09	.08	.02	.08^{-}	-.11	.04	.06	.07^{-}	**.16**	**.18**
local relevance	**.18**	**.14**	.08	.03	.01	.01^{-}	-.10	.02^{+}	.02	-.02	**.12**	**.13**
credibility	**.22**	.07	.06	.02^{-}	.05	-.01	**-.13**	.01^{+}	-.01	.00^{-}	.09	.08
emotional appeal	**.32**	**.22**	**.16**	**.12**$^{+}$	**.14**	.02^{-}	-.01	.10^{+}	-.01	.02	**.31**	**.25**
sufficiency	**.25**	**.18**	.06	.09^{-}	.00	.04^{-}	-.10	.03^{+}	.07	.06^{-}	**.15**	**.19**$^{+}$
IBM-AQR												
quality (WA)		**.08**		**.06**		**.01**		.00		**-.02**		**.08**$^{-}$
quality (MACE-P)		**.08**		**.05**		.01		.00		-.01		**.07**$^{-}$
stance		**.07**		.01		**-.03**		**.01**		**-.03**$^{+}$		**.04**
CORPS												
applause		**.02**		**.04**		**.07**		**.05**		**.01**		**.10**
laughter		**-.07**		**-.05**		**-.05**		**-.03**		**-.02**		**-.11**

Table 4: Spearman ρ between human annotations (HU) and multi-label BERT predictions (BM), respectively, and quality, stance and audience reactions. Bold values are statistically significant ($p < 0.05$). $^{+}/^{-}$: The correlation to the SBERT-Wiki predictions was considerably higher / lower (by at least 0.05).

data are as expected: a positive correlation for *applause* and a negative one for *laughter*, but again the effect is very weak.

Correlation with topic To control for topic effects, we computed the correlation between topic and argument quality and between topic and MF in the IBM-AQR. While we found no correlation between topic and argument quality, there was a weak correlation between some topics and specific MFs (see Table 5).

Topic	Moral Foundation	Spearman's ρ
The vow of celibacy should be abandoned	*Purity-Degradation*	.326
We should prohibit school prayer	*Purity-Degradation*	.298
We should ban targeted killing	*Care-Harm*	.237

Table 5: Topics with correlations to MFs greater than .2 according to multilabel-BERT predictions.

5 Discussion

A crucial issue for using moral values for argumentation analysis concerns the reliability of the (manual and automatic) annotations. Before we can reliably use moral values for the analysis of arguments, we need to ensure the quality of the annotations, as the low inter-annotator agreement for MF annotation casts doubt on the validity of the findings. While we expect that more extensive training and more detailed guidelines will increase IAA for human annotation at least slightly, we still think that due to the fuzziness of the concept of morality, high agreement scores are not very probable. Thus, we would like to propose a different approach to the annotation of moral foundations where we ground the annotations in lexical semantics. This approach has already been shown to improve IAA for a similarly difficult annotation task, namely the annotation of causal language (Dunietz et al., 2015). The authors created a lexical resource for terms that can trigger causality in text and instructed annotators to disambiguate instances of those terms in context, showing that their modularized, dictionary-based approach yields substantially increased IAA scores.

Inspired by their work, we propose to anchor MF annotations in lexical semantics, using an expanded version of the MFD as seed terms. The annotators will then be presented with instances of these terms and instructed to disambiguate them in context, and also to annotate a small, predefined set of semantic roles, such as *Betrayer, Harm_doer, Victim*, and so on. Setting up the annotation of moral values as a frame semantic annotation task has several advantages. First, the addition of semantic roles would make the annotations more informative by encoding the core participants of moral arguments, i.e., the target of the moral evaluation and the affected party. Second, anchoring the annotations in lexical semantics would make it easier to provide the annotators with precise guidelines. This might not only increase the consistency of the annotations but might also help to control for annotator bias. An open question, however, concerns coverage as it is not yet clear whether this approach would miss too many relevant expressions of moral values, given that it only captures explicitly stated moral evaluation but not implicit judgments. Whether the merits of our proposal outweigh its drawbacks needs to be explored in future work.

Being able to predict moral values in text reliably can open up new research avenues in argumentation. E.g., recent work in psychology has shown that moral values play an important role in debates on political and social issues (Feinberg and Willer, 2013; Voelkel and Feinberg, 2018; Feinberg and Willer, 2019). For example, Feinberg and Willer (2013) have shown that debates on environmental issues are often framed in terms of moral values such as *Care-Harm*, a moral foundation that is at the core of liberal belief systems, while conservatives, in contrast, seem to value all five MFs more similarly (Graham et al., 2009). This often results in highly polarized discussions, and Feinberg and Willer (2013) argue that reframing such issues in terms of moral values that explicitly address the opponents' belief system might have the potential to depolarize controversial debates and improve understanding between the camps by addressing the "moral empathy gap" (see Feinberg and Willer (2019) and references therein).

6 Conclusion and Future Work

In the paper, we evaluated different models for predicting moral sentiment in debates, based on Moral Foundations Theory. We then used our models to predict moral values in three argumentation datasets. We investigated whether we could find a correlation between morality and (i) argument quality, (ii) stance, and (iii) audience reactions for political speeches.

We found weak but significant correlations between general morality and argument quality in the ArgQuality data and a consistent positive correlation between moral sentiment and audience approval in CORPS as well as a negative correlation for moral sentiment and laughter. However, our study has several limitations that need to be addressed. First, the annotation experiment has been conducted by two annotators only, not allowing us to retrieve more reliable labels using the wisdom of the crowd. Also, the size of the test set is rather small, thus questioning the reliability of the results. Another problem is the low accuracy of the classifiers for the prediction of moral values. While results were substantially higher than the random baseline and an MFD-based baseline, we still expect a considerable amount of noise in the classifiers' predictions, which might impact the results of the correlation study. It is conceivable that cleaner predictions might increase the effect size of the observed correlations, which would be consistent with the slightly larger correlation coefficients found for the human annotations (HU). This, however, still needs to be confirmed.

The next steps should include the creation of larger test sets where the annotations have been validated by more than two annotators as well as the evaluation of semantically grounded approaches to coding moral values, to assess their reliability and validity. Another important task is the development of more accurate and robust classifiers for the prediction of moral sentiment.

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project ExpLAIN, Grant Number STU 266/14-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999), as well as within the SFB 884 on the Political Economy of Reforms at the University of Mannheim (projects B6 and C4).

References

Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. 2013. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Conference on Uncertainty in Artificial Intelligence*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196, Denver, Colorado, USA, June. Association for Computational Linguistics.

Matthew Feinberg and Robb Willer. 2013. The moral roots of environmental attitudes. *Psychological Science*, 24(1):56–62. PMID: 23228937.

Matthew Feinberg and Robb Willer. 2019. Moral reframing: A technique for effective and persuasive communication across political divides. *Social Psychology and Personality Compass*, pages 56–62.

Christiane Fellbaum. 2010. Princeton university: About wordnet.

Jeremy A. Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehgani. 2019. Moral Foundations Dictionary for Linguistic Analyses 2.0. Unpublished manuscript. Available from `https://osf.io/xakyw/`.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *The 10th International Conference on Language Resources and Evaluation*, LREC'16, pages 3730–3736, Paris, France, may. European Language Resources Association (ELRA).

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M. Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.

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

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, 47:55 – 130.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions. In Isabella Poggi, Francesca D'Errico, Laura Vincze, and Alessandro Vinciarelli, editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas, November. Association for Computational Linguistics.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Taher H. Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 0(0):0.

Kristen Johnson and Dan Goldwasser. 2018. Classification of Moral Foundations in Microblog Political Discourse. In *The 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, pages 720–730, July.

George Lakoff. 1997. *Moral Politics: What Conservatives Know That Liberals Don't*. University of Chicago Press.

Paul G. Lewis. 2019. Moral Foundations in the 2015-16 U.S. Presidential Primary Debates: The Positive and Negative Moral Vocabulary of Partisan Elites. *Social Sciences*, 8(233).

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring Background Knowledge to Improve Moral Value Prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.

Akiko Matsuo, Kazutoshi Sasahara, Yasuhiro Taguchi, and Minoru Karasawa. 2018. Development of the Japanese Moral Foundations Dictionary: Procedures and Applications. *CoRR*, abs/1804.00871.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the Measurement of Social Effects by Capturing Morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA, June. Association for Computational Linguistics.

Conny Roggeband and Rens Vliegenthart. 2007. Divergent framing: The public debate on migration in the Dutch parliament and media, 1995–2004. *West European Politics*, 3(30):524–548.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Julius Sim and Chris C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3):257–268, 03.

Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.

Hiroki Takikawa and Takuto Sakamoto. 2017. Moral Foundations of Political Discourse: Comparative Analysis of the Speech Records of the US Congress and the Japanese Diet. *CoRR*, abs/1704.06903.

Liling Tan. 2014. Pywsd: Python Implementations of Word Sense Disambiguation (WSD) Technologies [software]. https://github.com/alvations/pywsd.

Jan G. Voelkel and Matthew Feinberg. 2018. Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science*, 9(8):917–924. PMID: 30595808.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Ian Wood, John P. McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018. A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4654–4663, Hong Kong, China, November. Association for Computational Linguistics.