Enhancing the Measurement of Social Effects by Capturing Morality

Rezvaneh Rezapour, Saumil H. Shah, and **Jana Diesner**
School of Information Sciences
University of Illinois at Urbana-Champaign
{rezapou2, saumils2, jdiesner}@illinois.edu

Abstract

We investigate the relationship between basic principles of human morality and the expression of opinions in user-generated text data. We assume that people's backgrounds, culture, and values are associated with their perceptions and expressions of everyday topics, and that people's language use reflects these perceptions. While personal values and social effects are abstract and complex concepts, they have practical implications and are relevant for a wide range of NLP applications. To extract human values (in this paper, morality) and measure social effects (morality and stance), we empirically evaluate the usage of a morality lexicon that we expanded via a quality controlled, human in the loop process. As a result, we enhanced the Moral Foundations Dictionary in size (from 324 to 4,636 syntactically disambiguated entries) and scope. We used both lexica for feature-based and deep learning classification (SVM, RF, and LSTM) to test their usefulness for measuring social effects. We find that the enhancement of the original lexicon led to measurable improvements in prediction accuracy for the selected NLP tasks.

1 Introduction

User-generated text data are used in various fields to study, analyze, and extract people's culture, behavior, opinions, and emotions. The access and popularity of social media platforms such as Twitter attract individuals to participate in online discussions or share their points of view. Different beliefs and perspectives on social, political, economic, and other potentially controversial issues can lead to debates or conflicts among groups, and can result in arguments, abusive discussions, and segregated communities (Conover et al., 2011).

Given this type of behavior on online platforms, researchers have been investigating the relationship between basic principles of human values and the expression of opinions in user-generated text data by using (lexical) resources developed for this purpose and domain. This is done as part of stance analysis (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016), analysis of controversial topics (Addawood, Rezapour, Abdar, & Diesner, 2017), sentiment analysis (Wilson, Wiebe, & Hoffmann, 2005), and other standard NLP tasks. Following this line of research, in this paper, we operationalize and extract morality as a basic principle of human decision making and interaction guideline for people, e.g., when expressing themselves related to social or political topics. Our research is based on the assumption that people's backgrounds, cultures, and values affect their perception and expression of knowledge and beliefs about everyday topics. These personal idiosyncrasies and differences manifest themselves in people's social discourse and everyday use of language (Triandis, 1989), and can be helpful in analyzing or measuring people's positions or values regarding various social issues.

Concepts such as morality are challenging to measure as they require reliable operationalization and identification of regularities, and accounting for context and meaning (Bateson, 1972). To measure such concepts, we need to make sure that our results are - as much as possible - a reflection of the behavioral effect we want to study, not of the tools we use. The same is true for a wide range of social concepts that have been measured by applying lexicons to text data, such as opinion (Wiebe, Wilson, & Cardie, 2005), emotions

Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 35–45
Minneapolis, June 6, 2019. ©2019 Association for Computational Linguistics

(Munezero, Montero, Sutinen, & Pajunen, 2014), sentiment (Pang & Lee, 2008; Rezapour, Wang, Abdar, & Diesner, 2017), and culture (Van Holt, Johnson, Carley, Brinkley, & Diesner, 2013). Moreover, natural language text data are inherently ambiguous, and signals relevant for detecting personal characteristics and social effects are sparsely distributed across text data. Therefore, we can make the basic assumption that the reliable measurement of human behavior based on text data requires robust, reliable, and transparent tools to measure any effects in a credible fashion (Diesner, 2015). This paper contributes to this challenge by improving an off-the-shelve lexicon, known as the Morality Foundations Dictionary (MFD) (Graham et al., 2013; Graham, Haidt, & Nosek, 2009), and mitigating biases in measurement by expanding and validating the lexicon (enhanced MFD) by using multiple strategies and datasets. To achieve this goal, we performed a quality-controlled, semi-automated, and human-validated expansion of the original MFD (from 324 to 4,636 syntactically disambiguated entries) (discussed in Section 4). We then used the enhanced MFD as a feature for supervised learning to predict two social effects: (1) personal stance, and (2) individual value or morality (discussed in Section 5). To make a clear distinction between the two lexicons used in this paper, from this point, we refer to the original MFD as MFDO and to the enhanced lexicon as MFDE.

For predicting stance, we used semeval 2016 Stance detection benchmark dataset (Mohammad et al., 2016). For the second task, we leveraged the Baltimore protest benchmark dataset (Mooijman, Hoover, Lin, Ji, & Dehghani, 2017) created for predicting people's morality in tweets. The stance detection task is relevant to our assumption since individual differences in stance may relate to cultural differences. Therefore, we believe that the MFDE can be of assistance in improving the predictability of stance in user-generated texts. Regarding the second dataset, we found the Baltimore dataset relevant to our task since the dataset comes from the same domain, annotated on morality, and can show the usefulness of the MFDE lexicon.

The results of our prediction models show that using the MFDE as a feature outperformed prediction compared to MFDO. Using morality as a feature increased the performance of both classical feature-based (93%) and deep learning models (85.7%) in the majority of test cases. From that, we conclude that morality can be a useful feature for detecting social effects in text data. In addition, we observed that lexicon expansion is worthwhile as it improves prediction accuracy in the majority of experiments on both morality and stance prediction.

This study makes several contributions. First, we introduce and operationalize morality as a feature for NLP tasks, and show that incorporating this information can lead to measurable improvements in prediction accuracy of social effects such as stance. Second, we apply the morality lexicon not only for morality prediction, but also for stance prediction, and this out-of-domain test enhances the robustness of our findings. Third, we improve the accuracy and transparency of measuring morality based on text data, and provide a rigorous and reusable strategy for lexicon expansion and validation.

2 Literature Review

Moral Foundations Theory (MFT), introduced by Graham and Haidt, considers four sources of individual moral judgment: 1) innate features, 2) human learning, based on the cultural context in which people are embedded, 3) judgment based on situational intuition, and 4) pluralism of moral primitives (Graham et al., 2013; Graham et al., 2009; Haidt & Graham, 2007). Based on the MFT, the Moral Foundations Questionnaire (MFQ) was developed to facilitate measuring people's spontaneous morality (Graham et al., 2013). Such standardized questionnaires are often used by researchers to conceptualize morality and elicit information about moral reasoning from individuals in a lab or remote settings. Socio-demographic characteristics (e.g., age, gender) and personal characteristics (e.g., educational level, political orientation, religiosity) were often used to aggregate and compare the results of these questionnaires. While questionnaires and lab experiments provided valuable information, they entail some shortcomings such as high costs, limited scalability, mock-up setups, and reliability issues of self-reported data (Hofmann, Wisneski, Brandt, & Skitka, 2014).

Furthermore, alternative approaches like enhancement of a user study with neuro-physiological measures (Decety, Michalska, & Kinzler, 2012), AI-based simulations (Pereira & Saptawijaya, 2007), and extracting signals about morality from text data were used to address these shortcomings. In addition, text-mining techniques have been used to study user-generated, empirical data while eliminating issues with artificial lab settings and self-reported data.

The majority of prior studies that use NLP to study morality has focused on analyzing rhetorical aspects. Sagi and Dehghani (2014) used the MFDO to measure the moral loading of news data by analyzing articles about socio-political conflicts (World Trade Center before and after 1993 and 9/11 attacks, Ground-Zero Mosque and abortion) from the New York Times. In another study, Kaur and Sasahara (2016) leveraged a combination of the MFDO and latent semantic analysis to measure morality in tweets about different social issues, such as homosexuality and immigration. They found two dimensions, namely purity and care, to be dominant in conversations focused on immorality. Moral values have also been predicted using background knowledge and textual features. Lin and colleagues (2018) proposed a context-aware framework to aggregate external knowledge with text and improve morality prediction by 13.3% compared to the baseline. Garten and colleagues (2018) used a Distributed Dictionary Representations (DDR) approach to measure semantic similarity between dictionaries and text instead of using word counts. The DDR model was further used for predicting moral values of Twitter data related to Hurricane Sandy. Mooijman and colleagues (2017) evaluated the relation between online moral rhetoric and violent protests by applying Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) to a Baltimore Protests dataset. Dehghani and colleagues (2016) used the MFT to understand homophily, and found that people whose tweets are highly indicative of purity tend to be more like-minded. Finally, Fulgoni and colleagues (2016) leveraged the MFDO to analyze polarized debates in news sources. Their analysis showed different moral dimensions in liberals and conservatives conversations, where the former group favored care/harm and fairness, and the latter one focused on authority and loyalty.

Overall, a very few studies have extended the MFDO using variations of word embedding models and calculating the cosine similarities between moral foundation context vectors and word vectors (Kaur & Sasahara, 2016). Our work builds upon prior studies of MFDO expansions, but differs from them in that we evaluate the semi-automated and human-validated expansion of the original lexicon as a feature for NLP prediction problems. Our ultimate goal is not to improve morality prediction or stance detection (though we do, by a small margin), which are intensively studied problems in NLP. Instead, we aim to provide a rigorous strategy for lexicon expansion, and based on that a generally useful lexicon that can serve as a feature for a variety of information extraction and classification tasks. This can particularly be useful for people who want to use reliable resources.

3 Data

We used two public benchmark datasets that were previously annotated for morality (Baltimore) and stance. The Baltimore data[1] contains tweets related to the street violence that took place in Baltimore during the Freddie Gray protests (04/12/2015 to 05/08/2015). This dataset has been used to study if the rate of moral in tweets can assist in predicting violent protests (Mooijman et al., 2017). From 19 million tweets that were collected, the authors of the original paper removed those tweets for which the geolocation was not the same as the cities where protests related to the death of Freddie Gray took place. Next, they had human annotators code 5,000 tweets for moral content based on the MFT. The annotated tweets were then used to train a deep neural network-based model (RNN and LSTMs) to predict moral values from tweets; resulting in 89.01% accuracy. To get the dataset, we ran the tweet IDs through the Twitter API and were able to extract 3,793 of the tweets (around 75.8% of the original tweets) for which human labels were available.

The stance dataset was made available for SemEval 2016 (Mohammad et al., 2016). Using Twitter as a source, this dataset contains 4,870 tweets on six topics: abortion, atheism, climate change, feminism, Donald Trump, and Hillary Clinton. The tweets were hand-coded for stance, with the options being in favor, against, and none. The SemEval competition contained two tasks: Task A) was traditional supervised classification (on five topics mentioned above excluding Donald Trump), where 70% of the annotated data was used for training and the rest for testing. The highest

[1] https://psyarxiv.com/4bvyx/

accuracy (68.98%) was achieved by the baseline model, which used SVM and n-grams. Nineteen teams participated, and the best performing team achieved an overall accuracy (F-score) of 67.82% by using two RNN classifiers. Overall, about nine teams used some form of word embedding approaches, while some other teams leveraged publicly available lexicons (e.g., for sentiment, hashtags, and emotion), and Twitter specific features. For Task B, tweets on Donald Trump (a topic not used in Task A) were used. The highest F-score for Task B was 56.28% with nine teams participating. For our study, we combined the test and training sets from task A, and added the tweets on Donald Trump, resulting in a total of 4,870 tweets in our stance dataset.

4 Moral Foundations Lexicon Expansion

The Moral Foundation Theory (MFT) categorizes human behavior into five basic principles that characterize opposing values (virtues and vices) as shown in Table 1. To enable the measurement of this theory based on text data, the Moral Foundations Dictionary (MFD) was developed and published (Graham et al., 2013; Graham et al., 2009). In the original MFD, there is a sixth "miscellaneous" category, which is a collection of morally relevant words that were not yet mapped to any of the other categories. The MFDO associates 324 unique indicator terms (words) with the virtues and vices from the MFT. This lexical resource is highly valuable as it implements a theory. At the same time, it is limited in several ways: First, the number of entries is small and therefore might not capture all (variations of) terms indicative of morality in text data. This can lead to limited results, which may become part of our presumably valid knowledge about human morality. This problem can be mitigated through quality-controlled lexicon expansion as presented in this paper.

Second, we do not know based on what texts the MFDO was built, and even if we knew, these texts might be different from the ones to which researchers want to apply the MFDO. In NLP, this problem is known as domain adaptation. Several solutions to this problem have been developed (Daumé, 2007; Glorot, Bordes, & Bengio, 2011; Satpal & Sarawagi, 2007). Given that the MFT aims to measure basic principles of human behavior, one could aim to build a generally valid, i.e., robust and validated resources with broad term coverage, which can then be used as is or further be adapted to domains, contexts, and culture. We chose the second strategy as it results in an improved general resource for others (and us) to use, and present our solution to this problem in this paper.

Category	Virtue	Vice
Protecting versus hurting others	Care	Harm
Cooperation/ trust/ just versus cheating in interaction with objects and people	Fairness	Cheating
In-group commitment (to coalitions, teams, brands) versus leaving a group	Loyalty	Betrayal
Playing by the rules of a hierarchy versus challenging hierarchies	Authority	Subversion
Behavioral immune system versus spontaneous reaction	Purity	Degradation

Table 1: Moral foundations theory

In addition, the entries in the MFDO are not syntactically disambiguated, which can also limit the results, e.g., by capturing false positives. For example, one entry in the MFDO is *"safe,"* which represents the virtue of care. In a text, *"safe"* can occur as a noun, which is probably not the intended meaning, or as an adjective, which is more likely to be the intended meaning. This problem can be solved by adding the part of speech that represents the intended sense to each dictionary entry. We solve this problem as well.

The outlined limitations of the MFDO in terms of size, scope, and syntactic ambiguity can lead to flawed analysis results. We fixed these issues as described in the remainder of this section and tested the benefit of this work as described in the next section (Method).

To expand the lexicon, we first sorted the words from the "miscellaneous" category (which we named "general") into virtues and vices. Next, we manually annotated each lexicon entry with one or more best fitting parts of speech (POS). We then manually added variations of the original words and sense, such as grammatical inflections to the lexicon. All variations were added to the same category as the original root word. This expansion resulted in 1,085 words over 12 categories.

We then added synonyms, antonyms, and (direct) hypernyms of all original entries automatically by using WordNet (Fellbaum, 1998; Walenz & Didion, 2008); a word graph of broad scope and general applicability. To evaluate and adjust the new additions, we trained two human annotators to analyze every word entry for its POS and morality category assignment. Their initial intercoder-agreement was 65% (Kappa). After that, we went through all entries again, resolved annotation disagreements, and removed the words that the annotators found not suitable for any predefined category.

In the MFDO, some words occurred in multiple categories, which can confuse classifiers and make data analysis less robust. Therefore, we made the word to category assignment exclusive by assigning each redundant entry to only the best fitting category. To justify these assignments, we asked the human annotators to study each applicable term and choose the most suitable dimension for the words by considering their common meaning. Finally, we expanded nouns with their plural or singular form, adjectives with comparatives and superlative, and lemmatized the verbs (following the MPQA subjectivity lexicon (Wiebe et al., 2005)). Overall, our enhanced lexicon (MFDE) consists of 4,636 syntactically disambiguated, exhaustively expanded, and carefully pruned entries. Is this work worth the effort? To answer this question, we designed and ran experiments as described in the next sections. Our Enhanced Morality Lexicon can be accessed and downloaded at https://doi.org/10.13012/B2IDB-3805242_V1

5 Method

To analyze the impact of using the morality lexicons on predicting social effects, we built upon previous work in this domain. We assessed the performance of the lexicon and its expansion as features for both traditional feature-based and deep learning machine learning models. To test their impact on measuring social effects, we first created baseline models, and then added the original and enhanced MFD to the baseline to test if morality is a useful feature and if the learning with MFDE outperforms MFDO.

5.1 Data Preprocessing

Tweets are noisy in that they do not follow conventional spelling schemes, and therefore require extensive data cleaning and preprocessing. To prepare our datasets for analysis, we removed all URLs, mentions (usernames), hashtag symbols, punctuations, and numbers from the tweets. We then expanded contracted words by automatically converting them to their assumed intended form (e.g., “I’ve” to “I have”). Finally, we lowercased all words.

5.2 Classic Machine Learning

Figure 1 shows the overall experimental design used for this approach.

Feature Selection: We use morality words as additional attributes on top of the baseline models. We consider three types of counting to aggregate morality words per tweet: morality type count, morality dimension count, and morality polarity count.

Morality dimension count represents the number of words per tweet that match any of the five morality dimensions plus the general category, resulting in six attributes (each horizontal row in Table 1).

Morality type count represents the number of words per tweet that match words in the vice or virtue category of each morality dimension (each box in the last two columns of Table 1). Using the MFDO, this results in 11 additional attributes, and for MFDE in 12 (since we divided the general category into vice and virtue).

Morality polarity count represents the number of words per tweet that match any virtue or vice category regardless of the morality dimension (each of the last two columns in Table 1), resulting in two additional attributes.

We then test each counting approach with four feature sets: baseline (no morality feature), original morality, enhanced morality with POS, and enhanced morality without POS; all of which are explained next.

1) Baseline Model (BM): We replicated the baseline method from the SemEval competition from which we re-used the stance detection dataset. In the original SemEval competition, the best performing model was the baseline, which only used word level features, namely n-grams (Mohammad et al., 2016). To re-create that, we divided the dataset into its original sub-topics (feminism, climate change, atheism, Hillary Clinton, and abortion), and created one model for each sub-topic. We then replicated the unigram bag-of-words approach. To reduce the redundancy of the features, unlike in the original model, we removed stop words as well as words that appeared in less than 5 and more than 99% of the tweets. For the Baltimore dataset, we created a simple baseline

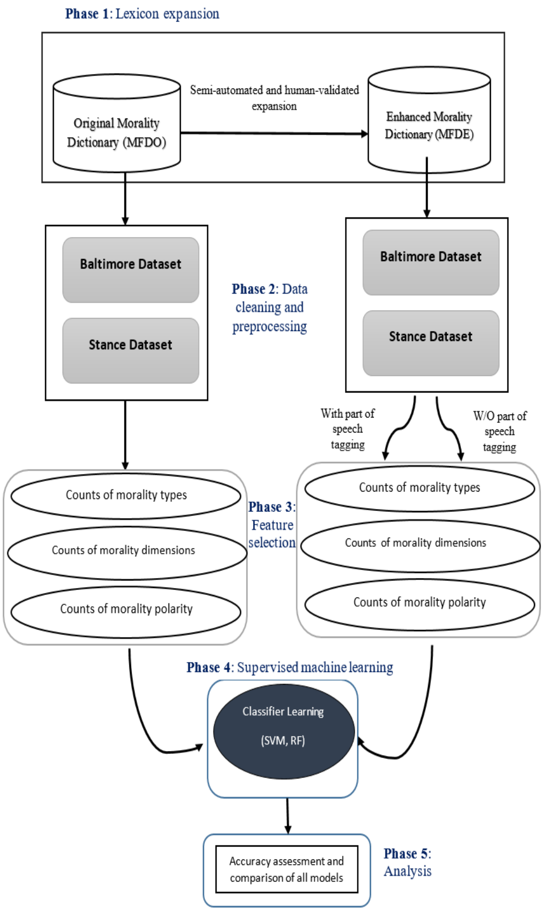
Phase 1: Lexicon expansion
Original Morality Dictionary (MFDO)
Semi-automated and human-validated expansion
Enhanced Morality Dictionary (MFDE)
Baltimore Dataset
Stance Dataset
Phase 2: Data cleaning and preprocessing
Baltimore Dataset
Stance Dataset
With part of speech tagging
W/O part of speech tagging
Counts of morality types
Counts of morality dimensions
Counts of morality polarity
Phase 3: Feature selection
Counts of morality types
Counts of morality dimensions
Counts of morality polarity
Phase 4: Supervised machine learning
Classifier Learning (SVM, RF)
Accuracy assessment and comparison of all models
Phase 5: Analysis

Figure 1: Experimental design and workflow of the classic machine learning approach

by extracting unigrams from the dataset and using the counts of words to create feature vectors.

We found that different numbers of tweets returned through the Twitter API as well as a lack of transparency for the original models, such as preprocessing steps and metrics, limited our ability to reproduce the original works.

2) Original Morality Model (OM): The MFDO consists of five dimensions that are further divided into virtue and vice and a sixth "miscellaneous" dimension. To aggregate the number of words per tweet, we used three types of counting as explained earlier. For the morality dimension, we added 6 attributes on top of the baseline (OM6), for the morality types, we added 11 attributes (OM11), and for morality polarity, we added two attributes to the baseline model (OM2).

3) Enhanced Morality Model with POS (EM): We used the Python NLTK library to tokenize the tweets and tag each token with a POS (Bird & Loper, 2004). We then used all matches between the texts and the MFDE if they agreed in POS as features. Finally, we aggregated the extracted words using the three counting methods explained above.

4) Enhanced Morality Model without POS (EMNP): To not only test the impact of dictionary expansion in size but also of word sense disambiguation based on syntax, we built a set of models where any word from tweets that matched the MFDE was considered regardless of its POS. This model results in a higher number of words in the BOW than the EM model since the grammatical agreement restriction was lifted from string matching. Again, we aggregated the extracted words using three count methods.

Classification: We used Support Vector Machine (SVM) and Random Forest (RF) as classification algorithms as implemented in the Python Scikitlearn package (Pedregosa et al., 2011).

For the stance dataset, we replicated the approach from the original SemEval task, i.e., we used a 70%-30% split for training and testing. For the Baltimore dataset, we conducted 5-fold cross-validation. To test the performance of our models, we (1) built the baseline model by using the full set of unigrams (BOW), (2) added attributes created from MFDO to the baseline model, and (3) added attributes created from MFDE with POS and (4) without POS to the baseline model for each of the two datasets. For each model, we tested the previously explained counting options (morality dimension, type, and polarity).

For assessing prediction accuracy, we used the standard metrics of overall accuracy, precision, recall, and F-score. Due to page limitation, we only report accuracy of the models (Table 2).

5.3 Deep Learning Models

We further investigated the usefulness of using lexicons using a recurrent neural network (RNN) with bidirectional long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). The advantage of LSTM compared to other RNNs is its ability to consider the whole context since it is capable of bridging long time lags between inputs. To implement the models, we used Keras (Chollet, 2018). For the stance dataset, we used a 70%-30% split for training and testing, and for the morality dataset, we used 5-fold cross-validation.

Baseline LSTM: To create the embedding layer, we leveraged the 200-dimensional word embedding from GloVe Twitter trained on two billion tweets (Pennington, Socher, & Manning, 2014). The embedding layer was followed by a Bidirectional LSTM of size 100, a hidden layer with Sigmoid activation function and an output layer with Softmax activation function. We further used Adam (Kingma & Ba, 2014) to optimize the parameters, and used cross-entropy as the loss function.

Enhanced LSTM with Morality Lexicon: To create the enhanced model, we first created the embedding layers of the lexicon words for (1) the MFDO (OM), (2) the MFDE with POS (EM), and (3) the MFDE without POS (EMNP). Moreover, we first found the words that intersected between the lexicon and datasets, and then created the embedding layers using the 200-dimensional GloVe Twitter (Pennington et al., 2014) without considering the morality dimensions, type, or polarity.

After that, we concatenated the output of the baseline Bidirectional LSTM (as explained above) with the embedding of the morality words to build three types of models: (1) OM, (2) EM, and (3) EMNP. After concatenating the LSTM output and lexicon embedding, we used a hidden layer with Sigmoid activation function and an output layer with Softmax activation function. We further used Adam (Kingma & Ba, 2014) to optimize the parameters, and used cross-entropy as the loss function.

One challenge in implementing neural network models is finding the best number of layers and settings (because there is no standard way of building the models). Since we are comparing different models, we found it challenging to choose a common set of numbers as the best hyperparameters, e.g., neurons, for both baseline and enhanced models. While we found one hidden layer to work best for our models, to increase transparency, we report the performance of our models with two sets of neuron sizes: 150 and 100. Table 3 shows the output of the LSTM models.

6 Results

Table 2 and 3 shows the result of predicting stance and morality. In both tables, the highest performance for each set of experiments (OM, EM and EMNP) is marked with bold text, and gray cells indicate the highest accuracy per model (per column).

The results for the ***classic machine learning models*** are shown in Table 2. For the Baltimore dataset (originally annotated for morality, last two columns in Table 2), using a simple set of basic unigram feature and classic machine learning models resulted in a baseline accuracy of 85.20% accuracy for SVM. Adding the simplest morality model (OM11) led to a small decrease (about 0.02%) with SVM. For the RF model, adding OM11 increased the performance by about 0.21%. Adding information about morality-relevant words in more sophisticated ways, (EMs and EMNPs) increased accuracy for both RF and SVM. As shown in Table 2, the best result for RF was achieved using EM2 (85.31%), and for SVM by using EM6 (85.71%).

Experiments		Stance Dataset												Baltimore	
		Abortion		Atheism		Climate		Clinton		Feminist		Trump			
		SVM	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*
Baseline	**BM**	66.42	62.5	69.54	64.54	61.76	**68.23**	60.81	60.13	58.94	60.7	51.17	45.07	85.20	83.91
Morality Types	**OM11**	66.42	62.5	**71.81**	65.0	**63.52**	**67.05**	61.14	57.77	**61.05**	**59.29**	50.7	49.29	85.18	84.12
	EM12	**67.85**	62.85	71.36	62.72	**63.52**	60.58	**64.18**	58.78	57.19	57.19	51.64	47.88	**85.60**	**84.73**
	EMNP12	66.07	**63.21**	71.36	**66.81**	62.35	62.94	62.38	**61.48**	58.94	**59.29**	**52.58**	**52.58**	85.31	84.12
Morality Dimension	**OM6**	**68.21**	**63.57**	70.45	**69.09**	**62.35**	64.7	59.79	58.1	59.29	60.7	51.17	46.94	85.31	**84.73**
	EM6	**68.21**	62.5	**71.36**	66.81	60.58	64.11	62.83	57.43	58.94	58.59	52.58	**53.99**	**85.71**	84.44
	EMNP6	**68.21**	62.5	70.45	60.0	60.0	**66.47**	**64.52**	**59.12**	**60.00**	**62.45**	**54.92**	50.7	85.55	84.10
Morality Polarity	**OM2**	67.14	63.21	69.09	**69.54**	**62.94**	65.29	62.83	57.09	58.24	56.84	52.58	**50.23**	85.31	84.99
	EM2	**67.85**	**64.28**	**72.27**	66.81	**62.94**	61.17	**63.17**	58.78	57.19	**61.05**	50.7	43.19	**85.60**	**85.31**
	EMNP2	67.14	63.92	71.81	64.54	61.17	**67.05**	**63.17**	**60.13**	**59.29**	56.49	**53.52**	49.29	85.49	84.84

Table 2: Result of predicting stance (first 12 columns) and morality (last two columns) with SVM and RF for stance and Baltimore datasets (Accuracy) (highest performance per set of experiments (OM, EM, and EMNP - each half column) in bold, highest accuracy per each model (each column) in gray)

#Neurons in Hidden Layer	Stance Dataset							Baltimore
	Experiments	**Abortion**	**Atheism**	**Climate**	**Clinton**	**Feminist**	**Trump**	
N = 150	**BM**	62.500	68.181	67.647	58.445	57.192	51.643	84.2391
	(1) OM	**68.214**	68.636	65.882	56.081	**57.894**	50.704	85.504
	(2) EM	67.500	72.272	**70.00**	**63.851**	**57.894**	50.234	**86.163**
	(3) EMNP	65.714	**73.181**	68.823	57.432	57.543	**54.929**	84.634
N = 100	**BM**	65.714	65.454	**70.588**	59.121	**58.596**	51.173	84.845
	(1) OM	64.642	66.363	69.411	**60.472**	56.842	51.643	85.900
	(2) EM	**67.142**	70.909	69.411	59.797	54.385	**53.521**	**86.612**
	(3) EMNP	64.642	**71.363**	67.647	56.756	58.245	49.765	83.580

Table 3: Result of predicting stance (first 7 columns) and morality (last column) with LSTM model for stance and Baltimore datasets (Accuracy) (highest performance per set of experiments (OM, EM, and EMNP – each half column) in bold, highest accuracy per each model (each column) in gray)

For the stance datasets, the results are shown in the first 12 columns of Table 2. Depending on the sub-topic, our baseline accuracy ranged from 45.07% (RF, Trump, stance hardest to predict) to 69.54% (SVM, atheism, stances easiest to predict). As observed for the Baltimore data, adding lexical morality features to stance increased accuracy over our baseline in all but one case (Climate, RF) cases.

The results for the ***LSTM model for both datasets*** are shown in Table 3. As mentioned before, we used two sets of neuron sizes for the hidden layer. For the Baltimore dataset, using the MFDE achieved better performance in both implemented models. The highest accuracy was obtained by the enhanced LSTM model using enhanced morality words (EM), 86.61% (N=100). For the stance dataset, adding morality embedding to the output of LSTM (baseline) resulted in outperforming the baseline in 83.33% of cases (10 out of 12).

Does using morality as a lexical feature improve prediction accuracy for the selected NLP tasks? Comparing the baseline to any models that include morality, we conclude that adding morality as a lexical feature increases accuracy in 13 out of 14 cases (93%) for feature-based learning (considering RF and SVM models for each topic) and in 12 out of 14 cases (85.7%) for deep learning (considering experiments with two sets of neurons for each topic). This finding suggests that using the morality as a feature is helpful for standard NLP tasks - and possibly other tasks as well, which would need to be explored in future work.

Does expanding the MFDO pay off? We find that for feature-based learning (Table 2), in 29 out of 42 cases (69.05%), the accuracy with any MFDE feature outperforms the models with MFDO features, in 21.43% of the cases, MFDO outperforms MFDE, and in 9.52% of the cases, both versions of the dictionary lead to equal results. For the LSTM, 9 out of 14 models (64.28%) had better performance when using MFDE, while 14.28% of models (2 models) worked better with MFDO (Table 3). From that, we conclude that lexicon expansion is worthwhile as it improves prediction accuracy in the majority of our experiments, especially for feature-based learning.

Does disambiguating word sense in the MFDO via POS pay off? Based on the results in Table 2 and 3, we found that syntactic disambiguating of lexicon entries leads to only minor quantitative improvements. We believe that the usefulness of POS tags can be further tested with other types of user-generated data that follow more conventional grammatical rules. In addition, beyond what we measured in this paper, this additional layer of information might further boost the quality of the data.

Based on the results of all implemented models, highlighted in Table 2 and 3, we found that using MFDE results in higher performance compared to other models (MFDO and BM).

7 Discussion and Conclusion

In this paper, we investigated the usefulness of leveraging morality as an NLP feature for predicting two selected social effects (morality and stance). In addition, we showed how investments in the quality and general nature of lexical auxiliary tools and the rigorous evaluation of these investments improve the predictability of these social effects, thereby reducing biases in algorithmic solutions. This work matters as personal values and social effects (which are often measured as the aggregation of personal values) are abstract and complex constructs, and their measurement requires researchers to find reliable and robust ways to operationalize these concepts. The validity of such research hinges on the trustworthiness of our methods for capturing these effects in digital traces of human behavior. Hence, our work is based on the assumption that people's personal values, which might be impacted by their cultural contexts, are reflected in their language use (Bateson, 1972; Milroy & Milroy, 1985; Triandis, 1989), and that we can capture these values in user-generated text data.

Enhancing lexicons is expensive, as it requires trained human coders to assess each entry and its meta-data (in our case, category assignment and part of speech). This might help to increase the reliability of social computing research, but does this effort make a difference for improving the accuracy of NLP tasks? In order to answer this question, we evaluated the usefulness of using no lexicon, a basic lexicon, and an enhanced lexicon for capturing morality in text data to measure two different social effects (morality and stance) based on public benchmark datasets. We found that using the lexicons we tested, namely the Moral Foundations Dictionary, does increase prediction accuracy in the majority of cases, especially when used for feature-based machine learning. Moreover, we found that the semi-automated and human-validated verification and advancement of this lexical resource led to measurable improvements in capturing social effects in text data.

Our work has several limitations. For deep learning models, while using the enhanced morality lexicon yielded better overall accuracy, we still need to investigate more parameters and settings to find the most robust models. We plan to investigate these settings in the future. Moreover, the benchmark data we used were too small for this purpose. In addition, we only worked with tweets, which is just one out of many types of user-generated text data. The robustness of our evaluation might be further improved by working with texts from other genres and of higher formality, such as debates, congressional speeches, product reviews, and news articles.

Acknowledgments

This research is sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We appreciate Jessica Schmiederer, Tiffany Lu, Craig Evans and Max McKittrick for their great help in expanding and evaluating the morality lexicon. We also thank Chieh-Li (Julian) Chin, Shubhanshu Mishra and Ly Dinh for their assistance.

References

Aseel Addawood, Rezvaneh Rezapour, Omid Abdar, and Jana Diesner. 2017. Telling apart tweets associated with controversial versus non-controversial topics. In Proceedings of the *Proceedings of the Second Workshop on NLP and Computational Social Science*, (pp. 32-41).

Gregory Bateson. 1972. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*: University of Chicago Press.

Steven Bird, and Edward Loper. 2004. NLTK: the natural language toolkit. In Proceedings of the *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, (pp. 31), Association for Computational Linguistics.

François Chollet. 2018. Keras: The python deep learning library. *Astrophysics Source Code Library*.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In Proceedings of the *Fifth international AAAI conference on weblogs and social media*.

Hal Daumé. 2007. Frustratingly easy domain adaptation. In Proceedings of the *45th Annual Meeting of the Association of Computational Linguistics (ACL)*, (pp. 256–263), (Vols. 45), Prague, Czech Republic.

Jean Decety, Kalina J Michalska, and Katherine D Kinzler. 2012. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. *Cerebral Cortex, 22*(1), 209-220.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, . . . Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General, 145*(3), 366.

Jana Diesner. 2015. Small decisions with big impact on data analytics. *Big Data & Society, 2*(2), 2053951715617185.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Dean Fulgoni, Jordan Carpenter, Lyle H Ungar, and Daniel Preotiuc-Pietro. 2016. An Empirical Exploration of Moral Foundations Theory in Partisan News Sources. In Proceedings of the *LREC*.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods, 50*(1), 344-361.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. *Proceedings of the 28th International Conference on Machine Learning*, 513-520.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55-130.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology, 96*(5), 1029-1046.

Jonathan Haidt, and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98-116.

Sepp Hochreiter, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Wilhelm Hofmann, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka. 2014. Morality in everyday life. *Science, 345*(6202), 1340-1343.

Rishemjit Kaur, and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on Twitter conversations. In Proceedings of the *Big Data (Big Data), 2016 IEEE International Conference on*, (pp. 2505-2512), IEEE.

Diederik P Kingma, and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In Proceedings of the *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (pp. 552-559), IEEE.

James Milroy, and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics, 21*(2), 339-384.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the *10th International Workshop on Semantic Evaluation (SemEval-2016)*, (pp. 31-41).

Marlon Mooijman, Joseph Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2017. When protests turn violent: The roles of moralization and moral convergence.

Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing, 5*(2), 101-111.

Bo Pang, and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1-2), 1-135.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, . . . Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825-2830.

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

Luís Moniz Pereira, and Ari Saptawijaya. 2007. Modelling morality with prospective logic *Progress in Artificial Intelligence* (pp. 99-111): Springer.

Rezvaneh Rezapour, Lufan Wang, Omid Abdar, and Jana Diesner. 2017. Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis. In Proceedings of the *2017 IEEE 11th*

International Conference on Semantic Computing (ICSC), (pp. 93-96), IEEE.

Eyal Sagi, and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Social science computer review, 32*(2), 132-144.

Sandeepkumar Satpal, and Sunita Sarawagi. 2007. Domain adaptation of conditional probability models via feature subsetting *Knowledge Discovery in Databases: PKDD 2007* (pp. 224-235). Springer-Verlag, Berlin: Springer.

Harry C Triandis. 1989. The self and social behavior in differing cultural contexts. *Psychological review, 96*, 506.

Tracy Van Holt, Jeffrey C Johnson, Kathleen M Carley, James Brinkley, and Jana Diesner. 2013. Rapid ethnographic assessment for cultural mapping. *Poetics, 41*(4), 366-383.

Brett Walenz, and John Didion. (2008). JWNL: Java WordNet library.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation, 39*(2-3), 165-210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.