MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction

Oscar Araque[a,*], Lorenzo Gatti[b], Kyriaki Kalimeri[c]

[a]*Intelligent Systems Group, Universidad Politécnica de Madrid, Avenida Complutense, 30, Madrid, Spain*
[b]*Human Media Interaction Lab, University of Twente, Enschede, The Netherlands*
[c]*Data Science Laboratory, ISI Foundation, Turin, Italy*

Abstract

Opinions and attitudes towards controversial social and political issues are hardly ever based on evidence alone. Moral values play a fundamental role in the decision-making process of how we perceive and interpret information. The Moral Foundations Dictionary (MFD) was developed to operationalize moral values in text. In this study, we present *MoralStrength*, a lexicon of approximately 1,000 lemmas, obtained as an extension of the Moral Foundations Dictionary, based on WordNet synsets. Moreover, for each lemma it provides with a crowdsourced numeric assessment of *Moral Valence*, indicating the strength with which a lemma is expressing the specific value. We evaluated the predictive potentials of this moral lexicon, defining three utilization approaches of increasing complexity, ranging from statistical properties of the lexicon to a deep learning approach of word embeddings based on semantic similarity. Logistic regression models trained on the features extracted from *MoralStrength*, significantly outperformed the current state-of-the-art, reaching an F1-score of 87.6% over the previous 62.4% (p-value< 0.01). Such findings pave the way for further research, allowing for an in-depth understanding of moral narratives in text for a wide range of social issues.

Keywords: Moral Foundations, moral values, lexicon, Twitter data, natural language processing, machine learning

1. Introduction

Social scientists, policymakers, and practitioners from diverse disciplines are paying ever-growing attention to digital data as they offer an alternative, complementary view of society. Psychological constructs are reflected in our digital behaviors [1, 2, 3]; with the burst of social media and online communication in

*Corresponding author
Email addresses: o.araque@upm.es (Oscar Araque), l.gatti@utwente.nl (Lorenzo Gatti), kkalimeri@acm.org (Kyriaki Kalimeri)

general, we now have access to linguistic content, timely, and at greater scale. Recent developments in the natural language processing domain, allow to understand better and model complex psychological processes such as emotion [4, 5], personality [6, 7], human values [8], and morality [9].

We place the focal point on the moral narratives, which influence the way we rationalize and take a stance upon a series of conflictual topics, like abortion, homosexuality, immigration, or religion [10]. Interestingly, they are related not only to politics [11] but also opinions about societal issues in general, including attitudes towards charitable donations [12, 13], climate change [14], poverty [15], vaccine hesitancy [16], or even violent protests [17] and terrorism [10].

We operationalise morality via the Moral Foundations Theory (MFT) [18], which expresses the psychological basis of morality, defining the following five foundations: *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, and *purity/degradation* (see [19, 20]). Even if in its infancy, MFT is the most well-established theory, both in the psychological and in the natural language processing domain, since it is the only theory that defines a clear taxonomy of values and a term dictionary, the Moral Foundations Dictionary (MFD, hereafter) [18]. Linguistic, cultural, and historical context reflect on language usage; Graham et al. [18], creators of the MFD, highlight the difficulty of creating such a resource. Among the most significant limitations of the MFD, we have: (i) a limited amount of lemmas and stem of words; (ii) "radical" lemmas rarely used in everyday language, for instance, "homologous" and "apostasy"; and (iii) an association with a moral bipolar scale, so-called vice and virtue, but without an indication of "strength".

Here, we address exactly these shortcomings; more precisely, we expanded the existing MFD using the WordNet lexical database [21]. Further, we provide a set of normative moral ratings for empirical assessment of moral narratives, going beyond the binary nature of the MFD. We present a machine learning framework exploring the potentials of *MoralStrength*, the proposed moral lexicon. More to employing feature extraction on the lexicon, we also present an approach based on semantic similarity, a metric computed between the text and the lexicons created ad-hoc, through the exploitation of embedding representations. The developed models are then employed to predict the moral narratives of a given text. In this way, we advance the current concept of the bipolar association of a given word with a moral dimension, to a state where we have not only a richer dictionary, but also an assessment of "moral valence" for each lemma.

We thoroughly evaluated the proposed lexicon on two datasets originating from the Twitter social media platform, annotated exactly as for their moral rhetoric according to the MFT foundations. Both datasets include tweets regarding critical social issues and have been previously employed in studies related to moral detection from text. Assessing the predictive power of our approach, we successfully improve the performance in predicting the moral narratives from the user-generated text for the current state-of-the-art methods. Further, we show that the combination of features exploiting *MoralStrength* with pure textual representations benefits the prediction performance. These findings pave the way

towards an in-depth understanding of the moral judgments, dispositions, and attitudes formation from social media data, which given their penetration to the population present an unprecedented opportunity to assess moral views in the wild.

Hence, we contribute to the research community and policy makers with a useful resource that can be employed for analyzing communication campaigns, or even nowcasting people's attitudes and opinions towards phenomena of significant social impact. Such knowledge is essential for policy-making specialists not only to understand how people perceive the information from mass media but also to design effective communication campaigns that appeal to people's values. Especially when it comes to critical issues, for instance, the assimilation of immigrants in the society, discrimination of minorities, or even attitudes towards vaccination, understanding personal narratives and viewpoints helps to forecast and prevent conflict.

2. Related Literature

Psychologists and social scientists have systematically analyzed text data to address their research questions. Iliev et al. [22] provided a review of the most popular approaches to automated text analysis from the perspective of social scientists. As the role of natural language processing (NLP) is steadily growing, the methodologies for text analysis for social sciences range from completely controlled, dictionary-based studies [8] to purely data-driven modeling of human values [6]. Here, we provide a brief review of the current literature on moral values assessment from textual data, starting from dictionary based studies to deep learning approaches.

The first vocabulary developed to assess the moral values from textual data was the Moral Foundations Dictionary (MFD) [18]. It was used together with the Linguistic Inquiry and Word Count (LIWC) software [1] to estimate moral traits and to investigate differences in moral concerns between different cultural groups. Clifford et al. [23] employed the MFD for performing a manual text analysis of 12 years of coverage in the New York Times focusing on a political debate in the US. Teernstra et al. [24] assessed the political debate regarding the "Grexit" from approximately 8,000 tweets. They compared the performance of using the raw data, bi-grams, and the MFD features in employing basic machine learning models, namely, Naive Bayes (NB) and Maximum Entropy (ME). They concluded that pure machine learning is preferable to dictionary approaches since it has similar prediction accuracy while using fewer assumptions. In this study, we follow a similar approach to [24]; however, we propose an expanded version of the MFD, including also the moral valence per lemma. Moreover, we employ logistic regression models to infer moral values from uni-grams combined with lexical features.

Dehghani et al. [25] examined the differences between liberal and conservative moral value systems using a hierarchical generative topic modeling technique based on Latent Dirichlet Allocation (LDA) [26] to enable the unsupervised detection of topics in their corpus of liberal and conservative weblogs.

They used small sets of words selected from the MFD as seeds to encourage the emergence of issues related to different moral concerns and examined similarities and differences in how such matters are expressed between these groups. Consistently with findings in moral psychology, they demonstrate that there are significant differences in how liberals and conservatives construct their moral belief systems. Sagi et al. [27] employed the same framework to study moral rhetoric in texts such as blogs and tweets, but also a specific case study of the U.S. Federal shutdown of 2013 [10] where they examined the role of morals in intra- and inter-community differences of political party retweets. In both works, they were based on the framework presented in [25], where LDA was employed to create a co-occurrence matrix on which the similarity between the texts and the vectors representing the different MFT moral traits was computed. In a similar approach, Kaur et al. [28] attempted to quantify the moral loadings of text, based on the Latent Semantic Analysis (LSA). They used a bag-of-words model, representing the entire corpus by a word-context matrix. Then they reduced its dimensionality obtaining low-dimensional word vectors, in which similar vectors represent similar meaning words. Our study is presenting a different approach since we do not use LSA representations, but rather pre-trained word embeddings models. Although pre-trained word embeddings do not contain domain-specific knowledge, they express language regularities encoded as offsets in the resulting vector space. The proposed representations based on the work of Araque et al. [29], exploit precisely the similarity between the analyzed text and a selection of words with moral content.

More recently, Garten et al. [30] employed the MFD to detect moral rhetoric in general, and more specifically, shifts in long political speeches over time. Then, based on psychological dictionaries and semantic similarity to quantify the presence of moral sentiment around a given topic, Garten et al. [9], proposed the Distributed Dictionary Representations (DDR) method. Showing promising results, DDR was also employed by Hoover et al. [13] to detect moral values in charitable giving, while later on, Garten et al. [31] extended the method, incorporating demographic embeddings into the language representations. Our approach is based on an expanded version of the MFD using WordNet synsets, with evaluated manual annotations regarding the moral valence of each lemma, that can be incorporated in computational frameworks.

In a study more similar to ours, attempting to predict moral values involved in Twitter posts automatically, Lin et al. [32] proposed a method that automatically acquires background knowledge to improve the moral value prediction, pointing out the difficulty of the task also for human experts. Based on the work of [32] and [18], [17] predicted the moral sentiment of the tweets. Their model consists of three layers, an embedding (lookup) layer, a recurrent neural network (RNN) with long short-term memory (LSTM) [33] and an output layer. The first layer converts words in an input tweet to a sequence of pre-trained word embeddings, the LSTM layer processes these embeddings and outputs a fixed-sized vector which encodes critical information for moral value prediction, while a vector representing the percentage of words that match each category in the Moral Foundations Dictionary [18] are concatenated with the LSTM feature

vector. Our approach is differentiating to this one since we employ word embeddings to compute the similarity between words rather than directly feeding them in a neural network architecture.

Core contribution of this study is the extended lexicon of moral lemmas with the respective moral valence. To exploit the properties and full potentials of the lexicon, we further suggest three different models of increasing complexity demonstrating the value of such resource. The proposed approaches range from feature engineering methods to a system which employs word embeddings of semantic similarity based on the work of Araque et al. [29].

3. Materials and Methods

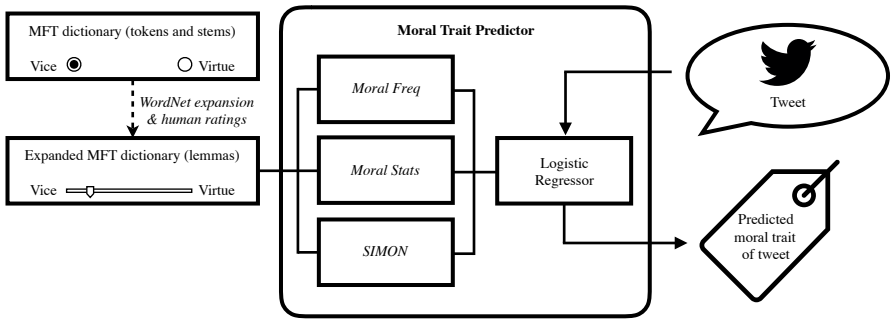

Figure 1: Overview of the process, from dictionary expansion to moral value prediction

3.1. Expansion of The Moral Foundations Dictionary (MFD)

The cornerstone of our study is the Moral Foundations Dictionary (MFD) [18] which was created to capture the moral rhetoric according to the five predefined dimensions defined by the Moral Foundations Theory (MFT) [34]. The original MFD[1] consists of lemmas and stems divided into "virtue" and "vice" [18] for each foundation according to their moral polarity. "Virtue" words are foundation-supporting words (e.g., *safe** and *shield* for Care "virtue"), whereas "vice" words are foundation-violating words (e.g., *kill* and *ravage* for Care "vice"). MFD [18] was meant to be used together with the Linguistic Inquiry and Word Count (LIWC) program [35], and thus contains either lemmas (158 entries such as *abandon*) or stems with a wild-card sign (166 listings), that LIWC analysis uses to match with all the forms of the base word; for instance, the entry *abuse** will match "abuse", "abuses", "abused", "abuser", "abusers", and so on. Due to the limited amount of lemmas and stems of words, often radical or

[1] Available at: `http://moralfoundations.org/othermaterials`

Moral Dimension	Virtue	Vice
Care/Harm	95 (16)	85 (35)
Fairness/Cheating	69 (26)	57 (18)
Loyalty/Betrayal	99 (29)	72 (23)
Authority/Subversion	160 (45)	101 (37)
Purity/Degradation	97 (35)	161 (55)
Total	520 (151)	476 (168)

Table 1: Corpus size after employing the WordNet resource to expand the MFD according to the official "virtue" and "vice" categories. The initial number of words contained in the MFD is shown in parenthesis.

rarely used in everyday language, for instance, "homologous" and "apostasy", the expansion of the existing dictionary is of essential importance.

Since we are interested in lemmas instead of stems, we initially expanded the original dictionary using the WordNet [21] synsets, maintaining the lemmas that shared the same initial part with stems in the MFD. The result of this first expansion was to obtain for each MFD entry, for instance, *traitor**, a series of lemmas, for instance, *traitor#n*, *traitorous#a*, *traitorously#r*, *traitorousness#n*[2].

We performed an initial preprocessing step on the obtained word corpus removing the forms that matched the search but clearly did not relate to a moral trait. For example, the stem *caste** not only matches *caste#n* and *caste_systems#n*, but also *caster#n* and *caster_sugar#n*, which are clearly not related to any moral foundation. This procedure was carried out manually, considering both the gloss for the lemma provided by WordNet and the moral trait that should be attributed to that word (e.g., while it could be argued that a statesman' name is an appropriate match for the Authority trait, the stem *church** relates to Purity, and thus we ignored *Churchill#n*).

Following the original classification, we divided the obtained word corpus (1,148 words) in "virtue" and "vice" lemmas resulting with 520 "virtues" and 476 "vices" while 152 were characterized as "general" morality words. These words can pertain to more than one traits; however, this is not common as the dataset consists of 442 unique "vice" words and 512 unique "virtue" words as shown in Table 1.

[2]The letter after the number sign # indicates the part of speech for that word, i.e., *#n* for nouns, *#a* for adjectives, *#r* for adverbs, and *#v* for verbs.

3.2. Moral Valence Annotation

Once the expanded dictionary was obtained, we used the Figure Eight[3] crowdsourcing platform to annotate each lemma with an association strength to the related moral trait. The goal here is twofold. On the one hand, we can use these annotations to determine if the terms extracted during the expansion process are still related to a moral trait. On the other hand, a lexicon with ratings could be useful for better dictionary-based approaches and is a first step in the direction of moral detectors that can rank sentences, instead of simply classifying them with a binary vice/virtue rating.

The expanded dictionary was annotated in terms of moral valence, but we also collected ratings of valence and arousal, following the definitions employed for the ANEW resource [21]. For our purpose, moral valence can be represented by a bipolar scale that, in aggregate, defines a continuous dimension from one moral extremity of the MFT to the other, e.g., from Care to Harm. Moral valence was operationalized in a 9-point Likert Scale, wherein if a word was ranked in the middle of the scale, it was semantically neutral to the specific moral dimension. The annotators were presented with the description of the moral trait and were asked to rate the relevance of the word to the specific foundation; if relevant, they were asked to rate its emotional valence, arousal and then its moral valence. Each experiment presented 20 different words to the annotator. The first time the annotators participated in the rating of a specific moral dimension (e.g., Care/Harm), after the experiment, they were asked to fill in the Moral Foundations Questionnaire [18] for the respective dimension. At least five annotators were recruited for each lemma.

The ratings of valence and arousal were included to ensure a minimum quality of the annotation. Since no existing resource annotates moral valence on a fine scale, we used the values of valence from the subset of words that appear both in our extended dictionary and in [36]. Annotators have always been presented 4 "gold" words among the 20 words they annotate, and the annotations of those who fail more than 1 gold word are discarded. A valid answer is one that lies within 1.5 standard deviations from the valence mean of [36], for each specific gold word.

3.3. Moral Lexicon Approaches

For the generated moral lexicon, we propose the following feature extraction approaches, which can be divided into those that solely exploit the semantic information of each word, and those who exploit the moral valence associated to the word. More specifically, we propose three lexicon utilization approaches: (i) counting, (ii) statistical summary, and (iii) word embedding similarity based representations. The two first approaches use both the words and their moral values, while the third one makes use solely of the selection of words, ignoring the associated numeric moral values.

[3]The Figure Eight Platform is available here: `https://www.figure-eight.com/`

Moral Freq. It consists of counting the number of words that express a specific moral dimension in a binary way. To decide if a specific word express a moral, we apply a simple rule: if the word has its moral value lower than a certain threshold, it does not convey that moral; if higher, the word does express that moral. Given the properties of the generated moral lexicon, the threshold is set at 5. We represent a given text with a 10-dimensional vector, which contains the corresponding normalized frequency counts, each for each moral extremity; for instance, *care/harm* are represented by two dimensions, one for *care* and other for *harm*.

Moral Stats. Given a specific text, we generate a statistical summary of the moral valence distribution of the contained words in the text. In the statistical summary, we included (i) the average, (ii) the standard deviation, (iii) the median, and (iv) the maximum value. As a result, the text is represented by a 20-dimensional vector.

SIMilarity-based sentiment projectiON (SIMON) (III). Finally, the third approach is known as SIMilarity-based sentiment projectiON (SIMON), described in [29]. This method was initially developed for sentiment analysis tasks, while here, we adapted it to moral valence assessment. *SIMON* uses a pre-trained word embedding model to compute the cosine similarity between the words of the analyzed text and a selection of domain related words, in our case a specific moral dimension. Projecting the analyzed text over the selection of words from *MoralStrength*, we result with a vector representation that encodes the similarity of the document to the specific moral dimension.

3.4. Evaluation Datasets

We evaluated our models on two Twitter datasets specifically collected to assess the moral values in user-generated content. Both datasets were employed in scientific studies assessing the moral narratives in the user-generated text according to the moral foundations' theory.

Hurricane Sandy Twitter Dataset. The first dataset we employed is presented in [32] and originally consisted of 4,191 tweets[4]. These Tweets contain hashtags relevant to the "Hurricane Sandy", a hurricane that caused significant damage to the Eastern seaboard of the United States in 2012, and they are annotated by experts indicating the presence or absence of each moral foundation dimension for each tweet. Moreover, annotations include a "non-moral" label, indicating that the specific text does not reflect any moral trait. Due to Twitter regulations, the original dataset could not be fully recovered, leaving us with only 3,853 messages. We further removed the retweets, keeping only the original messages, to avoid overfitting the data. In this way, the processed dataset consists of 3,478 instances.

Baltimore Protest Twitter Dataset. The second dataset is comprised of 18 million tweets [17]. These messages are related to the 2015 Baltimore Protests, which were motivated by the death of Freddie Gray. These data have been

[4]This dataset can be obtained from `https://osf.io/nzx3q/`.

Moral Foundation	Hurricane Sandy [32]	Baltimore Protest [17]
Care	217	8,040
Fairness	416	5,771
Loyalty	410	7,971
Authority	155	10,208
Purity	38	4,050
No moral	2,242	37,020
Total	3,478	73,060

Table 2: Statistics of the Twitter datasets employed in this study as a benchmark. The *Hurricane Sandy* dataset was annotated by human annotators, while, the *Baltimore Protest* was assigned the moral ratings employing a machine learning framework as described in the study [17]. Therefore, even if smaller in size, *Hurricane Sandy* is considered as the dataset of reference for this study.

automatically annotated by a machine learning classifier, as described in [17], and is freely available for research purposes. From this dataset, we randomly selected a subset of it, retaining 73,060 messages[5]. We opted for such selection to reduce computational complexity[6].

A shortcoming of the *Baltimore Protest* compared to the *Hurricane Sandy* dataset that is extremely important to consider is that the annotations included are not the result of human experts, but rather the output of a machine learning system. Thus, the annotations' quality is bounded by the error associated with the method proposed in the respective study [17]. In light of this, the *Hurricane Sandy* dataset remains the point of reference, since it does not entail intrinsic errors of any automatic system, but rather the opinion of human annotators. Table 2 reports the distribution of Tweets per moral dimension in the two datasets mentioned above. We note that among moral traits is not uniform, affecting the training of the algorithm as described in Section 4.2.

3.5. Experimental Design

After evaluating the crowdsourced annotations of *MoralStrength* (section 4.1), we suggest three main approaches on potential usage of the resource to analyze user-generated text (section 4.2), based both on feature engineering and word embeddings, as detailed in section 3.3. We evaluate the performance of the above approaches postulating the problem as a classification task.

[5]We report only the statistics of the subset used in this study.

[6]Namely, when performed over the whole 18 million instances, the experimental design (described in section 4) would take over 100 days for the entire computation and testing. For the restricted dataset, the complete evaluation takes no longer than 10 hours.

Hence, for the two datasets described above (see section 3.4), we train logistic regression models, employing 10-fold cross-validation, while reporting the F1-score as the evaluation metric per moral dimension. In our experimental design, we include a basic Bag-of-Words (unigram) model, which provides a standardized way of obtaining a baseline in the computational linguistics field. We built a series of logistic regression models; firstly, we assess the predictive power of the unigrams, *moral freq*, *moral stats*, and *simon* lexicons alone. Then, in a second step, we train logistic regression models concatenating the features extracted by the above approaches. We also combine the unigrams to the proposed lexicon approaches described above.

To directly compare our proposed framework with the current state-of-the-art approach of Lin et al. [32], we replicated their same configuration. Namely, we perform over-sampling on the original dataset to overcome the highly imbalanced nature of the benchmark data (see Section 3.4). After over-sampling on the *Hurricane Sandy* data, we resulted with an average number of training examples, $N = 6,128$, instead of the original dataset size, $N = 3,478$ (see Table 2).

Since over-sampling implies "artificial" data samples, we propose an alternative methodology; more specifically, we performed under-sampling, which also deals with the issue of unbalanced classes, however, in doing so, it randomly excludes data points of the most populated class. In this way, for the *Hurricane Sandy* we had $N = 824$ data points, while for the *Baltimore Protest* only $N = 24,026$ out of the original $N = 73,060$ remained. By reporting the score for both methods, we ensure the results are not biased by the technique used to address the class imbalance.

For all experiments, we report the performance in terms of F1-score, which is the metric also employed by Lin et al. [32], as well as the average F1-score over all moral dimensions. Moreover, to compare the improvement of the simplest model, which for this study we consider being the *Moral Freq* model, we employ the Friedman statistical test [37], which yields a ranking of the proposed method ordered by their performance. With the aim of obtaining further insights on the statistical significance of our obtained results for the baseline model, the Bonferroni-Dunn [37] post-hoc statistical test is performed with $\alpha = 0.05$.

4. Results and Discussion

4.1. Moral Valence Annotations

After collecting the moral valence ratings, we assessed the quality of the crowdsourced data with an intrinsic evaluation. However, since the only dictionary currently available for MFT has binary annotations (vice/virtue), a direct comparison with it is not informative enough.

Hence, we evaluate the quality by (i) calculating inter-annotator agreement for the moral valence ratings, (ii) calculating the correlation between valence scores and the normative lexicon of Warriner et al. [36], and (iii) comparing binarized moral valence ratings with the gold standard given by the MFD. The results for all these tests are reported in Table 3.

To assess inter-annotator agreement we calculated Gwet's agreement coefficient (AC2) [38]. We opted for this measure since other, more common measures (e.g., Cohen's Kappa) require the number of annotators per element to be constant, and this is not the case for our data. Moreover, Gwet's coefficient can be weighted, meaning that the coefficient score will be positively influenced by annotators expressing close ratings, and negatively influenced by scores that are far apart, a sensible feature for our dataset. Results for all the traits are in the "Moderate" to "Good" range (0.4-0.8), except for Fairness (which had "Poor" agreement, 0.17). While this is positive, it also indicates that the task is not trivial and that some words might be hard to rate. The lower agreement of Fairness led us to inspect the agreements for all traits manually, and we discovered that some annotators were particularly inaccurate. It was thus decided to discard some annotators, despite their ability to complete the crowdsourced experiment without failing the control questions. In particular, for the Authority trait, the annotator with the worst agreement was removed, improving the original AC2 of 0.41 to 0.42. For Loyalty, the answer of one annotator was lost due to programmatic error (the result for one word is outside the range specified by the Likert scale) and was removed from the dataset (no effect on the agreement). In the case of Fairness, we intervened more drastically and removed 5 annotators, plus 1 non-valid answer. The 5 discarded annotators were chosen due to them having a poor agreement with other annotators, and to inconsistent ratings (i.e., they gave the same rating to antonyms that have opposite traits in the MFD gold standard, such as "honest" and "dishonest"). The inter-annotator agreement for valence ranges between 0.61 and 0.72, thus falling in the "Good" category for the set of words of every moral trait. This indicates, in general, that annotating valence is easier and less controversial than rating moral traits.

We also compared the aggregated values of valence ratings (i.e., the mean of all valence annotations for a word) with the gold scores provided by [36]. In this case, we report the results of the Pearson correlation, which ranges from 0.79 to 0.95, indicating once again that the crowdsourced annotation is of good quality, and that differences between annotators are within the acceptable range.

Finally, to be able to compare with the only gold standard for moral foundations, i.e., the Moral Foundations Dictionary, we binarized the aggregated annotations and excluded those whose average is 5 (the center of the Likert scale, meaning that the word is neither positive nor negative[7]). This way, we could calculate Cohen's kappa coefficient by comparing to the vice/virtue ratings of the MFD for the subset of words that exists in both datasets. The lowest agreement is for Authority, but also, in this case, the 0.78 value suggests that the annotations are generally reliable and quite in line with the original MFD. It is perhaps worth noting that the agreement of Fairness is quite good (0.84), despite the lower inter-annotator agreement of the collected ratings. This might

[7]While it would be sensible to consider neutral a range instead of a single value, e.g., excluding everything in the interval 4.5-5.5, we wanted to avoid removing more words from the comparison.

Moral trait	Inter-annotator	Warr correlation	MFD agreement
Authority	0.42	0.84	0.78
Care	0.65	0.95	0.91
Fairness	0.34	0.88	0.84
Loyalty	0.59	0.91	0.84
Purity	0.56	0.79	0.92

Table 3: Measures of the quality of the collected ratings. The first column is the inter-annotator agreement for each moral dimension via Gwet's gamma with quadratic weighting metric. The second column is the correlation of the aggregate valence ratings and the gold standard of [36]. The last column is the agreement of the aggregate ratings (binarized) and the original Moral Foundations Dictionary, using Cohen's Kappa.

indicate that, while the aggregate ratings are reliable (i.e., they fall in the correct side of the morality spectrum), there is a relatively high individual variation regarding where the words of that dimension should be placed.

4.2. Evaluation of the Lexicon Usage

In this section, we present an assessment of the predictive power of the various approaches exploiting *MoralStrength* on the benchmark datasets described above.

Evaluation on Hurricane Sandy. Initially, we report the performance of the predictive model employing unigrams. Such a model represents an objective baseline, an assessment of the difficulty of the task itself. Then, we report the performance of the logistic regression models inferring on a specific set of features extracted from the expanded moral lexicon, i.e., *Moral Freq, Moral Stats*, and *SIMON*, followed models inferring on aggregations of the above lexicons. Table 4 describes the exact combinations of lexicons on which each model is trained as well as the obtained results for the over-sampling approach on the *Hurricane Sandy* dataset. For all experiments, we report the Friedman statistical test [37], which yields a ranking of the proposed methods ordered by their performance. With the aim of obtaining further insights on the statistical significance of our obtained results with respect to the baseline model, the Bonferroni-Dunn [37] post-hoc statistical test is performed. Note that in this study, for the statistical significance test, we employed the *Moral Freq* model as a baseline model, and not the unigram one, since it is the one that infers on the simplest generated lexicon.

Across all moral dimensions, the model inferring on the aggregated unigram and SIMON features emerges as the best performing approach; with a statistically significant improvement of the average F1-score - 87.6 over 62.4 reported by Lin et al. [32].

Interestingly, the highest score is obtained for "purity", which was reported being the most challenging moral dimension in the work of Lin et al. [32]. Ex-

amining each moral dimension separately, we note that our results are also consistently higher than the unigram model. The models that stand out are (i) unigrams + *SIMON* for fairness, loyalty, and purity, (ii) unigrams + *SIMON* + *Moral Freq* for care, purity, and neutral text, while (iii) unigrams + *SIMON* + *Moral Freq* + *Moral Stats* is the best performing models for authority and purity.

As described in section 3.5, we also applied an under-sampling technique for dealing with the unbalanced training data, a methodology which does not generate artificial data samples. Table 5 reports the results of this evaluation. Following this approach, the results vary with respect to oversampling (see Table 4), while the average overall performance improves (88.2% against 87.6% F1-score). At the same time, the best performing models, of four out of five moral dimensions, yield notably higher results with the respective model of the oversampling technique despite the smaller sample size.

Interestingly, the importance of the statistical features regarding the moral valence of lemmas, exploited in the *Moral Freq* and *Moral Stats* Lexicons, is more pronounced for all moral traits, with respect to the oversampling technique. More precisely, the model inferring from unigrams together with the *Moral Stats* model, has a better performance in fairness, loyalty, and purity, while for care, the best performing model is the SIMON combined to the *Moral Freq* and *Moral Stats* Lexicons. Observing the obtained results, we conclude that combining lexicon-driven representations which take into consideration the moral valence, together with pure textual information (for instance, the unigrams), allows for a more robust and semantically meaningful representation.

Despite the differences in the proposed approaches, we further compare our approach to the study presented by Garten et al. [30], that also predicted the moral foundations on the *Hurricane Sandy* dataset. Their better model achieved 49.6% F1-Score, which is remarkably lower than the 88.2% reported here.

Evaluation on Baltimore Protest. We extended the evaluation, employing the *Baltimore Protest*, which is considerably larger in comparison to the *Hurricane Sandy* one. The *Baltimore Protest* dataset is presented in the study of Mooijman et al. [17]. Since their approach is not directly comparable to ours, we perform the validation process only on training data obtained with the under-sampling technique. Note that ground-truth here are annotations obtained via a neural network model, and hence, our model's evaluation is bounded to the performance of the system presented in Mooijman et al. [17]. We note that the overall performance for all moral dimensions is lower to the respective performances on *Hurricane Sandy* (see Table 5), with the sole exception the case of Loyalty, with the average F1-score to be 87.2%. Similar as before, for all moral dimensions, the best performing model infers on the features obtained from the unigrams; in three out of five dimensions the best model is the one inferring on the combined with the *SIMON* word embeddings, while for the remaining two, from the combination with *Moral Freq* and *Moral Stats* Lexicons. Of course, unigrams combined with the SIMON representation are more effective with more massive datasets, since the resulting vector representations are larger and more generic; however, the importance of the information entailed in the moral va-

Approach	**C/H**	**F/C**	**L/B**	**A/S**	**P/D**	**NM**	**Avg.**	**Rank**
State of the art: Lin et al. [32]	82.3	70.7	50.3	69.3	37.4	64.2	62.4	12.3
unigrams	74.0	76.9	76.5	80.7	94.1	77.2	79.9	11.9
Moral Freq	61.4	58.2	61.9	56.0	62.1	63.4	60.5	14.0
Moral Stats	62.8	57.2	58.8	52.7	64.1	63.3	59.8	14.3
SIMON	79.6	82.3	77.1	86.0	98.1	84.2	84.5	6.6*
SIMON + *Moral Freq*	79.2	82.5	77.2	83.8	98.2	83.9	84.1	6.7*
SIMON + *Moral Stats*	79.2	82.2	77.0	84.0	98.2	83.9	84.1	7.6
SIMON + *Moral Freq* + *Moral Stats*	79.6	82.5	77.1	84.0	98.2	83.8	84.2	6.8*
unigrams + *Moral Freq*	75.3	77.7	77.2	81.2	95.5	77.8	80.8	9.8
unigrams + *Moral Stats*	73.5	77.6	76.7	81.3	95.7	77.9	80.5	10.8
unigrams + *Moral Freq* + *Moral Stats*	74.0	78.2	77.1	81.7	95.9	77.9	80.8	9.3
unigrams + *SIMON*	84.6	**85.6**	**81.2**	90.0	**98.9**	85.5	**87.6**	2.1*†
unigrams + *SIMON* + *Moral Freq*	**85.1**	85.2	80.8	89.5	**98.9**	**85.6**	87.5	2.4*†
unigrams + *SIMON* + *Moral Stats*	84.9	85.4	80.4	90.0	98.8	85.2	87.5	3.3*†
unigrams + *SIMON* + *Moral Freq* + *Moral Stats*	85.0	85.4	80.8	**90.2**	**98.9**	85.2	**87.6**	2.3*†

Table 4: F1-Score of the proposed methods using over-sampling over *Hurricane Sandy* ([32]). C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '$\star$' and '$\dagger$' mark that the approach significantly outperforms the *Moral Freq* and [32] baselines, respectively. The model with the lowest rank is the one that outperforms the rest.

lence still emerges.

To conclude, we observe that the performance trends are maintained; a unigram model is a robust approach, and adding information from *MoralStrength* improves the prediction performance. We believe that exploratory analysis will be useful for the ever-increasing studies on moral foundations since it presents a variety of approaches on how the moral lexicon we propose can be employed for the prediction of moral narratives from a text.

5. Conclusions

There is an ever-increasing interest in moral values understanding since they reflect our perception, attitudes, and opinion formation on critical societal issues. Moral values are expressed in user-generated content, and primarily through text. With the burst of social media data, emerges a unique opportunity of observing such behaviors in scale and as they happen. Recent developments

Approach	**C/H**	**F/C**	**L/B**	**A/S**	**P/D**	**NM**	**Avg.**	**Rank**
unigrams	87.3	90.6	83.5	88.3	89.4	82.0	86.8	9.1
Moral Freq	62.0	59.4	63.2	56.8	70.1	64.1	62.6	13.5
Moral Stats	70.6	66.8	70.3	66.6	66.7	68.6	68.3	13.2
SIMON	85.7	85.2	84.1	90.6	69.3	85.3	83.4	9.2
SIMON + Moral Freq	87.3	83.9	83.7	90.9	69.3	84.4	83.3	9.6
SIMON + Moral Stats	**90.1**	85.0	83.9	90.6	81.4	84.5	85.9	6.7*
SIMON + Moral Freq + Moral Stats	**90.1**	85.4	84.2	91.0	78.8	84.4	85.6	6.3*
unigrams + *Moral Freq*	88.2	**92.1**	**85.7**	87.7	**92.1**	83.0	88.1	5.3*
unigrams + *Moral Stats*	89.6	91.5	85.6	88.0	90.7	83.3	88.1	5.4*
unigrams + *Moral Freq + Moral Stats*	89.8	91.7	85.5	88.0	90.7	83.6	**88.2**	5.0*
unigrams + *SIMON*	86.9	91.3	85.3	90.3	74.9	86.8	85.9	7.2
unigrams + *SIMON + Moral Freq*	88.5	91.5	84.8	91.3	77.6	87.1	86.8	5.1*
unigrams + *SIMON + Moral Stats*	88.0	90.7	85.3	91.9	80.2	**87.9**	87.3	4.9*
unigrams + *SIMON + Moral Freq + Moral Stats*	88.0	90.9	85.3	**92.3**	80.2	**87.9**	87.4	4.6*

Table 5: F1-Score of the proposed methods using under-sampling over *Hurricane Sandy* ([32]). C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '*' marks that the approach statistically outperforms the *Moral Freq.* baseline. The model with the lowest rank is the one that outperforms the rest.

Approach	**C/H**	**F/C**	**L/B**	**A/S**	**P/D**	**NM**	**Avg.**	**Rank**
unigrams	83.5	86.9	87.2	84.0	89.3	82.8	85.6	6.9*
Moral Freq	47.9	43.3	46.1	41.6	45.5	41.9	44.4	13.7
Moral Stats	48.3	42.0	45.4	42.2	51.9	43.0	45.5	13.3
SIMON	79.8	84.0	81.5	81.2	89.4	80.6	82.7	8.7
SIMON + Moral Freq	79.6	79.1	81.1	80.0	86.8	80.3	81.1	11.4
SIMON + Moral Stats	80.1	79.2	81.4	80.0	86.8	80.3	81.3	10.7
SIMON + Moral Freq + Moral Stats	80.2	79.3	81.4	80.0	86.4	80.3	81.3	10.6
unigrams + *Moral Freq*	84.0	86.9	**87.9**	84.1	88.5	83.0	85.7	5.2*
unigrams + *Moral Stats*	85.1	86.2	87.7	83.7	88.0	83.0	85.6	5.4*
unigrams + *Moral Freq + Moral Stats*	**84.9**	86.1	87.7	83.7	88.0	**83.0**	85.6	5.8*
unigrams + *SIMON*	84.6	**89.8**	87.3	**85.1**	**90.1**	86.2	**87.2**	2.5*
unigrams + *SIMON + Moral Freq*	84.1	89.1	87.3	84.1	89.5	86.2	86.7	4.6*
unigrams + *SIMON + Moral Stats*	84.3	89.5	87.6	84.5	89.6	86.2	87.0	3.3*
unigrams + *SIMON + Moral Freq + Moral Stats*	84.4	89.5	87.6	84.5	89.6	86.2	87.0	3.1*

Table 6: F1-Score of the proposed methods using under-sampling over *Baltimore Protest* ([17]). C/H: Care/Harm, F/C: Fairness/Cheating, L/B: Loyalty/Betrayal, A/S: Authority/Subversion, P/D: Purity/Degradation, NM: Non-moral, Avg.: Average. '$\star$' marks that the approach statistically outperforms the *Moral Freq.* baseline. The model with the lowest rank is the one that outperforms the rest.

in the computational linguistics domain, allow us to analyze automatically such data obtaining useful insights.

Operationalizing morality via the Moral Foundations Theory (MFT) [18], we propose a linguistic resource, *MoralStrength*, that aims at improving the only currently available dictionary, i.e., the MFD. More specifically, we contribute with a moral lexicon containing (i) a large number of lemmas, (ii) less radical and more frequently used lemmas, hence, improving its usability, and (iii) finally, containing a metric of moral valence for each lemma. *MoralStrength* contains approximately five times more lemmas than the MFD, while at the same time providing with a moral valence, i.e., a quantitative assessment to characterize the lemmas' relationship with each moral dimension.

To explore the potentials of the moral lexicon in predicting the moral narrative in an unseen text, we generated three representations employing a series of feature extraction techniques, including normalized lemmas frequencies, statistical features, and finally, semantic similarity based on word embeddings. We evaluated the machine learning framework on two benchmark datasets from the Twitter platform, the only available resources of linguistic data explicitly annotated for their moral content.

Interestingly, all our models improve the prediction performance with respect to the current state-of-the-art for all moral dimensions. The most prominent approaches - as indicated by the Friedman ranking - combine pure textual (e.g., unigrams) with lexicon-based representations (e.g., the *Moral Freq*, the *Moral Stats*, and the SIMON). Hence, we argue that moral lexicon can be successfully employed for moral values classification from a given text since when this information is considered, the models yield higher performance.

This study paves the way for further advancements in the moral text analysis, which is indeed an exciting field of study, both from the computational linguistics and the social sciences points of view. From a linguistic perspective, it would be interesting to explore how specific knowledge could be encoded in domain-oriented word vectors, allowing for the development of complex learning methods. Moreover, the word embedding representation based on moral similarity could be enhanced with the obtained assessments of moral valence, or even combined with sentiment features from the analyzed text. As for the social sciences, there are numerous issues where detecting the morals narrative can significantly improve our understanding of the peoples' dispositions in, for instance, controversial social phenomena or news reporting biases.

References

References

[1] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, Journal of language and social psychology 29 (1) (2010) 24–54.

[2] R. L. Boyd, J. W. Pennebaker, Language-based personality: a new approach to personality in a digital world, Current Opinion in Behavioral Sciences 18 (2017) 63–68.

[3] K. Kalimeri, M. G. Beiró, M. Delfino, R. Raleigh, C. Cattuto, Predicting demographics, moral foundations, and human values from digital behaviours, Computers in Human Behavior 92 (2019) 428–445.

[4] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, Cambridge, UK, 2015.

[5] C. Strapparava, R. Mihalcea, Learning to identify emotions in text, in: Proceedings of the 2008 ACM symposium on Applied computing (SAC), Fortaleza, Ceara, Brazil, 2008, pp. 1556–1560.

[6] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PloS one 8 (9) (2013) e73791.

[7] T. Yarkoni, Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers, Journal of Research in Personality 44 (3) (2010) 363–373.

[8] R. L. Boyd, S. R. Wilson, J. W. Pennebaker, M. Kosinski, D. J. Stillwell, R. Mihalcea, Values in words: Using language to evaluate and understand personal values, in: Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM), Oxford, UK, 2015, pp. 31–40.

[9] J. Garten, J. Hoover, K. M. Johnson, R. Boghrati, C. Iskiwitch, M. Dehghani, Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis, Behavior Research Methods 50 (1) (2018) 344–361.

[10] E. Sagi, M. Dehghani, Moral rhetoric in Twitter: A case study of the US Federal Shutdown of 2013, in: Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci), Vol. 36, 2014, pp. 1347–1352.

[11] A. Miles, S. Vaisey, Morality and politics: Comparing alternate theories, Social Science Research 53 (2015) 252 – 269.

[12] K. P. Winterich, Y. Zhang, V. Mittal, How political identity and charity positioning increase donations: Insights from moral foundations theory, International Journal of Research in Marketing 29 (4) (2012) 346 – 354.

[13] J. Hoover, K. Johnson, R. Boghrati, J. Graham, M. Dehghani, Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation, Collabra: Psychology 4 (1) (2018).

[14] C. Wolsko, H. Ariceaga, J. Seiden, Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors, Journal of Experimental Social Psychology 65 (2016) 7 – 19.

[15] M. Low, M. G. L. Wui, Moral foundations and attitudes towards the poor, Current Psychology 35 (2016) 650–656.

[16] A. B. Amin, R. A. Bednarczyk, C. E. Ray, K. J. Melchiori, J. Graham, J. R. Huntsinger, S. B. Omer, Association of moral values with vaccine hesitancy, Nature Human Behaviour 1 (2017) 873–880.

[17] M. Mooijman, J. Hoover, Y. Lin, H. Ji, M. Dehghani, Moralization in social networks and the emergence of violence during protests, Nature Human Behaviour 2 (6) (2018) 389–396.

[18] J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations, Journal of Personality and Social Psychology 96 (5) (2009) 1029.

[19] J. Haidt, J. Graham, When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize, Social Justice Research 20 (1) (2007) 98–116.

[20] J. Haidt, C. Joseph, Intuitive ethics: How innately prepared intuitions generate culturally variable virtues, Daedalus 133 (4) (2004) 55–66.

[21] G. A. Miller, Wordnet: A lexical database for english, Communications of the ACM 38 (11) (1995) 39–41.

[22] R. Iliev, M. Dehghani, E. Sagi, Automated text analysis in psychology: Methods, applications, and future developments, Language and Cognition 7 (2) (2015) 265–290.

[23] S. Clifford, J. Jerit, How words do the work of politics: Moral foundations theory and the debate over stem cell research, The Journal of Politics 75 (3) (2013) 659–671.

[24] L. Teernstra, P. van der Putten, L. Noordegraaf-Eelens, F. Verbeek, The morality machine: tracking moral values in tweets, in: Proceedings of the 15th International Symposium on Intelligent Data Analysis (IDA), Stockholm, Sweden, 2016, pp. 26–37.

[25] M. Dehghani, K. Sagae, S. Sachdeva, J. Gratch, Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "Ground Zero Mosque", Journal of Information Technology & Politics 11 (1) (2014) 1–14.

[26] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning research 3 (2003) 993–1022.

[27] E. Sagi, M. Dehghani, Measuring moral rhetoric in text, Social science computer review 32 (2014) 132–144.

[28] R. Kaur, K. Sasahara, Quantifying moral foundations from various topics on Twitter conversations, in: Proceedings of the 2016 IEEE International Conference on Big Data (BigData), Washington D.C., USA, 2016, pp. 2505–2512.

[29] O. Araque, G. Zhu, C. A. Iglesias, A semantic similarity-based perspective of affect lexicons for sentiment analysis, Knowledge-Based Systems 165 (2018) 346–359.

[30] J. Garten, R. Boghrati, J. Hoover, K. M. Johnson, M. Dehghani, Morality between the lines: Detecting moral sentiment in text, in: Proceedings of the IJCAI 2016 Workshop on Computational Modeling of Attitudes (WCMA), New York, NY, USA, 2016.

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

[32] Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, H. Ji, 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), Barcelona, Spain, 2018, pp. 552–559.

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[34] J. Graham, B. a. Nosek, J. Haidt, R. Iyer, S. Koleva, P. H. Ditto, Mapping the moral domain, Journal of Personality and Social Psychology 101 (2) (2011) 366–85.

[35] J. W. Pennebaker, The secret life of pronouns, New Scientist 211 (2828) (2011) 42–45.

[36] A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, Behavior Research Methods 45 (4) (2013) 1191–1207.

[37] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (Jan) (2006) 1–30.

[38] K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters, Advanced Analytics, LLC, Gaithersburg, MD, USA, 2014.