Knowledge Graphs meet Moral Values

Ioana Hulpuș[1], Jonathan Kobbe[1], Heiner Stuckenschmidt[1], Graeme Hirst[2]
[1]University of Mannheim
[2]Department of Computer Science, University of Toronto
{ioana,jonathan,heiner}@informatik.uni-mannheim.de,
gh@cs.toronto.edu

Abstract

Operationalizing morality is crucial for understanding multiple aspects of society that have moral values at their core – such as riots, mobilizing movements, public debates, etc. Moral Foundations Theory (MFT) has become one of the most adopted theories of morality partly due to its accompanying lexicon, the Moral Foundation Dictionary (MFD), which offers a base for computationally dealing with morality. In this work, we exploit the MFD in a novel direction by investigating how well moral values are captured by KGs. We explore three widely used KGs, and provide concept-level analogues for the MFD. Furthermore, we propose several Personalized PageRank variations in order to score all the concepts and entities in the KGs with respect to their relevance to the different moral values. Our promising results help to progress the operationalization of morality in both NLP and KG communities.

1 Introduction

Many of the choices that we make in daily life, such as political stance or position in debates on ideological topics, are influenced by our moral values (Sagi and Dehghani, 2014; Wolsko et al., 2016; Amin et al., 2017). Besides, moral values and moral judgments are central to decision making and cultural cohesion (Dehghani et al., 2016). The last years have seen an increasing interest in operationalizing the concept of morality as defined by psychologists, particularly from the NLP, social media, and communication communities, into an effort of extracting the latent moral dimension of texts.

Tweets (Garten et al., 2016; Araque et al., 2020), newspaper articles (Bowman et al., 2014), as well as scientific articles (Clifford and Jerit, 2013) or religious sermons (Graham et al., 2009) have been targeted for moral analysis. Lately, this line of research has been widely motivated by the rise of social media campaigns such as #BlackLivesMatter and #MeToo, which have a very strong moral load. People take different stances with respect to such matters depending on their understanding and hierarchy of moral values, and can lead to clashes of visions even within the same culture. The general assumption is that the words used in discussions on such topics reveal the moral values of the discussants.

One of the most widely adopted theories of morality is the Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Graham et al., 2013). MFT proposes at least five moral foundations, each one consisting on the one side of virtues and on the other side of vices: (1) **the care/harm foundation** which deals with the sensitivity towards the suffering of others; (2) **the fairness/cheating foundation** covering aspects of reciprocity and motivations to be fair; (3) **the loyalty / betrayal foundation** covering aspects of in-group cooperation, and the intuition of being loyal to one's group; (4) **the authority/subversion foundation** which is related to the innate intuition of endorsing hierarchies that we find just; (5) **the purity / degradation foundation** which deals with our innate drive of preferring cleanliness of body and soul over hedonism.

The Moral Foundations Dictionary (MFD) (Graham et al., 2009) has been proposed as a lexicon to guide the assessment of the moral foundations of the MFT in texts. It consists of a set of words and lemmas for each vice and virtue of each foundation and has become an essential resource for operationalizing moral values.

Nevertheless, being a word-level lexicon, this resource comes with several limitations. First, the natural ambiguity of language means that some

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 71–80
Barcelona, Spain (Online), December 12–13, 2020

of the words and lemmas provided can have other meanings than the ones related to the moral foundations. For instance, the stem *subver** covers the ambiguous word *subversion*, which besides the meaning related to the *authority / subversion* foundation, it is used with a completely different meaning in software development.

Second, it contains a limited set of lemmas and words, particularly focusing on those whose main meaning is the one related to the corresponding moral foundation. This, on the one side, means that more ambiguous synonyms are not contained, and on the other side, it means that very specific rarely used words are contained. These effects lead to a high precision with poor recall, which is not necessarily the preferred strategy in many scenarios.

Moreover, it only contains uni-grams, and the entries are associated with either vice or virtue, without a score for the strength of the association. However, this lexicon has been widely exploited lately, and several approaches have been proposed to overcome these weaknesses, for instance, by extending it or by its projection into continuous spaces.

In this work, we investigate a new direction, that of projecting the MFD lexicon on knowledge graphs (KGs) with the purpose of scoring all entities and concepts therein with respect to their relevance for each moral foundation. We envision multiple benefits from this endeavor. First, it overcomes the ambiguity and incompleteness limitations. Second, entities and concepts in KGs oftentimes strongly relate to moral values. For example, *Rebecca Reichmann Tavares* is a UN diplomat promoting race relations and human rights, a position highly related to the Fairness/Cheating moral foundation. Similarly, the concept *History of Human Rights* is also highly related to the same moral foundation. Such concepts and entities are not part of the MFD, but their mentioning can help the detection of moral foundations expressed by the texts.

Third, the usefulness of KGs for many tasks resides in the fact that they provide factual knowledge such as relations between entities. Still, an important drawback of current KGs is their weak representation of common sense knowledge. For example, the fact that the concept of "crime" is generally bad and undesirable cannot be derived from current KGs although such common sense knowledge is crucial, for example, for understanding arguments: To understand that an argument claiming that *racism leads to crime* is an argument against racism, it is necessary to understand that crime is *bad* or a *vice*. In this work, we propose a means of adding a moral dimension to KGs, and hence extend them with common-sense "intuition" of morality[1].

2 Related Work

The main target of this work is to investigate how the moral foundations characterized by the MFD are captured in KGs, with the purpose of scoring each entity and concept in the KGs with respect to their relevance to the moral foundations. Therefore, we are particularly interested in the way previous literature uses the MFD.

Some works (Graham et al., 2009; Clifford and Jerit, 2013; Teernstra et al., 2016; Lin et al., 2018) simply use MFD counts either on their own or as features in supervised classifiers for determining the moral values expressed by text. Works that try to overcome the issues related to the simple counts of lexicon hits embed the moral values in continuous spaces. Dehghani et al. (2016) and Kaur and Sasahara (2016) generate vectors for words based on a Latent Semantic Analysis (LSA) (Deerwester et al., 1990) methodology. Then, for each moral value, a vector in the same space is obtained by adding up all the vectors of the modeled lexicon words. Garten et al. (2016), Nokhiz and Li (2017) and Xie et al. (2019) use Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) models to embed the words of the lexicon, and then either aggregate these vectors to produce a vector representation of each moral value, or they keep each of these vectors separate and use different strategies for determining the moral values expressed by novel embedded texts (such as k-nearest neighbor).

Other works that go into a direction more related to ours extend the MFD lexicon by using WordNet (Rezapour et al., 2019; Araque et al., 2020), and then manually curate the results with human annotators (Rezapour et al., 2019) or extend the annotations with values for valence and arousal (Araque et al., 2020). However, the methods differ from our approach as the results of these works are still word-level lexicons, while our moral value relevance annotation is done at a synset rather than

[1] All resources created in this work, as well as all relevance scores for the KGs, are available at https://github.com/dwslab/Morality-in-Knowledge-Graphs.

word level. For example, MoralStrength (Araque et al., 2020) is built by manually expanding the MFD with new words from the same WordNet synsets.

The only work that acknowledges that entities from Wikipedia/DBpedia also carry a moral load is that of Lin et al. (2018). They use Wikipedia abstracts and DBpedia properties in order to generate features of the entities mentioned in text and use these features together with the textual features in order to classify the texts based on their moral values. However, they do not provide a moral value score of the specified entities, and they do not look into linking the MFD to the knowledge base.

With respect to WordNet, this work is related to another research direction, that of automatically creating sense-level lexicons starting from word-level lexicons. Two commonly used such lexicons are SentiWordNet (Baccianella et al., 2010) and +/–EffectWordNet (Choi and Wiebe, 2014). The first one is a sense-level lexicon providing WordNet synsets with prior sentiment polarity annotations. +/–EffectWordNet provides annotations for synsets that express negative or positive effects over entities. Our work of annotating WordNet synsets as well as other KG concepts with scores reflecting their relevance to moral values is complementary to the previously mentioned lexicons, as they bring in the dimension of morality.

In our work, we investigate how moral values are captured in KGs such as DBpedia, WordNet, and ConceptNet. We use the same intuition across all three KGs: if terms in the MFD can be linked to corresponding concepts in the KG, then the semantic relations contained in the KGs can be exploited to score all the other concepts with respect to their relevance to the moral values. These relevance scores can subsequently be used, for example, for expanding the lexicon. Alternatively, they can be used directly as features in applications aiming at classifying texts based on moral values.

Another important motivation for our work is that KGs have the benefit of providing structured knowledge, but most of the time, this knowledge is factual (e.g., DBpedia) or lexical (e.g., WordNet). Most KGs, ConceptNet being one exception, lack in common sense knowledge. And knowledge such as what is generally accepted as morally good or bad is also missing in ConceptNet. This work is a step in the direction of enriching KGs with such common-sense knowledge.

3 Approach

The core idea of our approach is to map the MFD lexicon to a KG, and subsequently score all entities in the KG with respect to their relevance to moral values. We are taking a layered approach to the moral foundations, as we are interested in scoring the relevance on three levels: (1) the **moral trait level**, in which we are interested in the relevance of concepts with respect to each moral trait (virtue / vice). Hence, each concept obtains 10 scores, two (virtue and vice) for each moral foundation. (2) **the moral foundation level** that scores each concept with respect to the five moral foundations; and (3) the **the moral polarity level** that scores each concept with respect to its relevance to vices and virtues.

As a measure of relevance in KGs, we use Personalized PageRank (Haveliwala, 2003)(PPR). Our work consists of two main steps: first, we manually link the MFD entries to entries in the KGs. Then, we use these KG entries as seeds for running PPR. We now describe both steps in more detail.

3.1 Linking the MFD to KGs

We manually link each entry in the lexicon to the corresponding concept(s) in the KGs. This step involves disambiguation judgments. Also, multiple concepts can be linked to the same lexicon entry, as long as they are related to the corresponding moral value.

Linking to WordNet 3.1 (WN) The concepts in WN which we focus on are the synsets. As each synset has a specific meaning, we aim to only link to concepts that are truly relevant for the respective moral foundation. For example, the word *fair* occurs as a synonym for *just* which is relevant for the moral foundation of *fairness/Cheating*, but also as a *gathering of producers to promote business*. Thus, we manually decided for each synset, which contains an entry of the MFD lexicon, whether it relates to the moral foundation or not.

Linking to ConceptNet (CN) In CN, the concepts are only disambiguated with respect to their part of speech. As this kind of disambiguation is not needed for linking the entries and in order to obtain a graph that is less sparse, we collapsed all concepts that only differ in their part of speech. Further, we remove so-called *External Concepts*, which are links to other resources such as WordNet as well as all isolated nodes. For classifying the

specific moral traits and the moral polarity, we remove such relations between concepts that express a semantic difference: *Antonym, DistinctFrom, NotCapableOf, NotDesires, NotHasProperty*. For linking to ConceptNet, we check which single word concepts match the lexicon entries and manually verify these links to avoid nonsense (i.e., *Churchill* matches *church**, but we do not expect it to be an intended cue word for *purity/sanctity*). As this way of linking often includes inflected forms of words (i.e. *care** also includes *cares*, *cared*, *caring* and even *careth*), we further ignore *DerivedFrom* and *FormOf* relations in our leave-one-out evaluation.

Linking to DBpedia For linking the lexicon entries to concepts in DBpedia, we use the following process: we check in Wikipedia for articles whose name is the lexicon entry. If such an article exists, it is related to the moral value, and it is not a disambiguation page, then we add the link to the sense lexicon. If the formed URL is redirected, we check if the redirected article is related to the moral value. If so, then we add it to the sense lexicon. If an article or a redirected page does not exist, but a disambiguation page, then we check each disambiguation article listed in the disambiguation page and select the ones that are related to the moral trait. This is the case, for instance, for the term *shelter* with Wikipedia disambiguation page `www.wikipedia.org/wiki/Shelter`. This page provides many disambiguation options for the term, including *Shelter_(building)*, *Animal_shelter*, *Homeless_shelter* which we add to the lexicon, but also others that we deem unrelated to the intended meaning of the MFD, for example, multiple locations such as *Port_Shelter* - a harbor in Hong Kong, multiple films, albums, singles, novels, video games, and others that we do not add to the lexicon. In Wikipedia, it is often the case that an article is the default for a particular word, but that a Wikipedia disambiguation page also exists for the same word. If that is the case, we add the default article if it is related to the moral trait, and also check the disambiguation page, and manually select from all the disambiguation articles the ones that are related to the moral trait. We then map all the collected Wikipedia articles to their corresponding DBpedia resources.

The exact number of concepts that we obtain in this process for each moral trait and each KG are shown in Table 1. We use these concepts as seeds for computing relevance scores of all the concepts in the KG, as described in the following section.

Moral value	MFD	WN	CN	DBpedia
A-virtue	43	77	391	44
A-vice	27	74	285	28
C-virtue	16	66	361	53
C-vice	35	121	304	31
F-virtue	26	40	145	36
F-vice	18	46	180	24
L-virtue	28	48	260	34
L-vice	22	39	203	15
P-virtue	34	22	269	19
P-vice	46	52	482	40
General	29	82	343	26
Total	324	667	3223	350

Table 1: Number of concepts per moral trait.

3.2 Personalized PageRank for Moral Value Relevance Scoring

We use Personalized PageRank as a measure of relevance in KGs. Relations in KGs are typed and directed. However, the relations are semantic. Therefore for each relation, one can consider that another relation in the opposite direction also exists. For example, for each *occupationOf* relation, the *hasOccupation* relation can be defined, pointing in the opposite direction. Since PPR is working on directed networks, but we want the random walker to be able to also follow incoming edges, we add for each relation with type t in the KGs, an additional, opposite relation whose type we set to *inv_t*.

We investigate 3 ways of computing the link probabilities in PPR: **Uniform (U)** which is the standard PPR, disregarding the edge types; **Type-Uniform (TU)**: the random walker chooses uniformly at random one of the available edge types; Then, for the chosen type, it chooses uniformly at random one of the edges with that particular type; **Type exclusivity (TE)**: the random walker first chooses an edge type uniformly at random. Then, among the edges of that type, it chooses which one to take according to the exclusivity of the edge. Exclusivity (Hulpuş et al., 2015) is a measure of relation importance that provides higher scores to relation types with low cardinality for both source and target nodes.

On the **moral trait level**, we consider each of the 10 classes of traits in the MFD individually, excluding the *general* class whose entries are not split according to their polarity. For each trait, we run a PPR process where the teleport probability is distributed uniformly to the seeds of the trait. Consequently, each concept in the KG receives 10

	Uniform			Type-Uniform			Type-Exclusivity		
Class	WN	CN	DBpedia	WN	CN	DBpedia	WN	CN	DBpedia
A-virtue	.58	.87	.42	.55	.86	.39	.55	.86	.46
A-vice	.50	.72	.39	.50	.72	.43	.50	.72	.32
C-virtue	.77	.98	.46	.77	.96	.40	.76	.96	.46
C-vice	.66	.86	.32	.64	.87	.42	.65	.86	.35
F-virtue	.73	.95	.44	.68	.94	.50	.68	.93	.53
F-vice	.54	.96	.79	.52	.96	.71	.52	.96	.67
L-virtue	.50	.85	.26	.52	.85	.26	.54	.95	.29
L-vice	.49	.96	.13	.54	.96	.13	.54	.93	.13
P-virtue	.59	.91	.53	.64	.92	.53	.64	.88	.47
P-vice	.46	.88	.52	.46	.88	.55	.46	.85	.55
Overall	.59	.89	.43	.59	.89	.44	.59	.88	.44

Table 2: Prediction accuracy for all moral traits, KGs and PR methods

scores for each PPR method.

On the **moral foundation level**, we score all concepts with respect to their relevance to the five moral foundations. We create the set of seeds for each foundation by merging the corresponding sets of vice and virtue seeds. One PPR process is run for each foundation, as well as on the *general morality* class, therefore each concept in the KG receives 6 scores for each PPR method. As previously, the teleport probability of the PPR process is shared uniformly by the seeds.

On the **moral polarity level**, we create the set of seeds for each of the two classes (vices and virtues) by merging the vice seeds and the virtue seeds of all foundations, respectively. Therefore, we provide each concept with two scores for each PPR method.

4 Experiments and Discussion

We evaluate the prediction on each level independently. Similarly to Xie et al. (2019), we use a leave-one-out evaluation. Specifically, for each seed concept of a class, we run an additional PPR process for that class when the targeted seed concept is left out of the teleport vector. Then, we check the relevance score the targeted seed concept obtains in this PPR process and compare it to its relevance scores for the other classes.

To measure the accuracy of the prediction, we take a very straightforward approach and consider a hit when the targeted seed concept achieves the maximum relevance score for the class to which it belongs when it is left out. Therefore, we compute the accuracy for a class c as the percent of seed concepts of c that obtained the maximum score across all classes for class c, when their score for class c is computed as their PPR score when left out of the seed set.

4.1 Results of Moral Trait Prediction

Table 2 shows the results obtained for each PPR version, per trait as well as overall, for each of the three KGs. Since there are 10 classes, a random baseline assignment would obtain, on average, a .10 score. Therefore, all methods manage to perform substantially better than random on all KGs.

Regarding the methods, we observe that the scores differ between the classes, particularly for DBpedia. For example, the Uniform PPR achieves a .79 score for class Fairness-vice on DBpedia, the Type Uniform PPR achieves .71, and the Type-Exclusivity .67. The class Authority-vice is also handled very differently by the three methods on DBpedia. On the other KGs, the different PPR methods do not show significant differences, with their performances being most similar on ConceptNet. Indeed, DBpedia provides many relation types with varied cardinality, so it is not surprising that methods that treat relation types differently obtain significantly different results on this KG. However, interestingly, the overall results of all the three methods are very similar, including on DBpedia.

With respect to KGs, ConceptNet is in a strong lead over both WordNet and DBpedia, with all methods obtaining scores between .72 and .98 on the individual moral traits. All methods perform worse on WordNet with .59 overall scores. These values are similar to the highest scores obtained by Xie et al. (2019) when running their leave-one-out classification of moral traits, specifically where the MFD entries are embedded using the Google N-grams corpus (Lin et al., 2012) and the classification is done with a Centroid model.

DBpedia captures the moral traits the worst among the three KGs, with a wide range of values across different classes and an overall score of .43 or .44 depending on the method. This perfor-

	Predicted class									
True class	A-virtue	A-vice	C-virtue	C-vice	F-virtue	F-vice	L-virtue	L-vice	P-virtue	P-vice
A-virtue	.58*	.08	.05	.01	.03	.01	.04	.09	**.10**	.00
A-vice	.07	.50*	.01	.09	.05	.03	.03	**.15**	.03	.04
C-virtue	.03	.00	.77*	.05	**.06**	.02	.05	.02	.02	.00
C-vice	.03	**.07**	.02	.66*	.02	.03	.02	.07	.02	.05
F-virtue	.00	.05	**.08**	.00	.73*	.00	.05	.05	.05	.00
F-vice	.04	.02	.02	.07	.02	.54*	.07	**.15**	.04	.02
L-virtue	.08	.04	.06	.08	.04	.06	.50*	**.10**	.02	.00
L-vice	.00	.10	.05	.08	.05	.05	**.13**	.49*	.05	.00
P-virtue	**.18**	.00	.05	.00	.05	.00	.00	.00	.59*	.14
P-vice	.00	.06	.02	.12	.04	.12	.02	.02	**.15**	.46*

Table 3: Confusion matrix of moral trend prediction on WordNet, for Uniform PPR

	Predicted class									
True class	A-virtue	A-vice	C-virtue	C-vice	F-virtue	F-vice	L-virtue	L-vice	P-virtue	P-vice
A-virtue	.46*	**.14**	.05	.00	.09	.05	.05	.05	.09	.02
A-vice	.11	.32*	.00	**.18**	.04	.11	.04	.11	.04	.07
C-virtue	**.12**	.02	.46*	.08	.08	.04	.00	.08	.04	.08
C-vice	.03	.06	**.16**	.35*	.03	.03	.03	**.16**	.00	.13
F-virtue	**.11**	.00	.05	.03	.53*	.08	.05	.03	.08	.03
F-vice	**.12**	.00	.00	.00	.08	.67*	.08	.00	.08	.00
L-virtue	.12	.03	.09	.00	.12	**.26**	.29*	.03	.06	.00
L-vice	.00	**.20**	.13	**.20**	.13	.13	.00	.13*	.00	.07
P-virtue	.10	.05	.00	.05	.05	.00	.00	.00	.47*	**.26**
P-vice	.00	.1	.02	.00	.05	.02	.02	.02	**.20**	.55*

Table 4: Confusion matrix for the prediction of moral traits on DBpedia, for Type-Exclusivity PPR

mance is still better than all the methods proposed by Xie et al. (2019) when training the MFD lexicon entry embeddings on the COHA corpus[2].

Regarding the different moral traits, several scores stand out, particularly the almost random performance of all methods on DBpedia for the Loyalty-vice class. As seen in Table 1, we managed to identify only 15 DBpedia concepts for this class. Among them, many are also present in the set of concepts of other moral traits. For instance, `dbpedia.org/resource/Apostasy` also belongs to the Authority-vice and to the Purity-vice classes, while three concepts related to MFD entry *abandon* are also part of the seed concepts of Care-vice. Among the MFD entries that only occur in this class, for many, we did not find a corresponding concept in DBpedia, for instance, *imposter, jilt*, miscreant, renegate*. The confusion matrix shown in Table 4 reveals that more entries of the Loyalty-vice class achieve higher relevance with respect to the Authority-vice and Care-vice traits rather than with respect to the Loyalty vice trait. Also on WordNet, Loyalty-vice prediction is relatively often mistaken for Authority-vice and Care-vice, as seen in the confusion matrix of Table 3.

As seen in Tables 3 and 4, in both WordNet and DBpedia, vices are usually confused with vices of another foundation, and virtues with virtues of other foundations, and when that is not the case, then a vice is confused with a virtue of the same dimension and the other way round. An exception is the prediction of Loyalty-virtue on DBpedia, which is often mistaken for Fairness-Vice. This is likely due to their shared MFD lexicon entry *segregation*, for which we encounter 7 concepts in DBpedia.

4.2 Results of Moral Foundation Prediction

Table 5 shows the results obtained on the moral dimension level, for each PPR version, per class as well as overall, for each of the three KGs. Since we consider 6 classes, a random baseline achieves on average .17, therefore again, all methods on all KGs significantly outperform this trivial baseline.

The Overall results come to reinforce our conclusion from the previous analysis that the treatment of relation types in the PPR process is only beneficial for DBpedia. As expected, the results of this prediction are better than of the moral-trait prediction, and this improvement is particularly strong on DBpedia.

Among the foundations, the Authority/Subversion foundation achieves under average scores on all methods on all KGs, while the Care/Harm foundation is correctly predicted more often than average by all methods on all KGs. The

[2] `https://www.english-corpora.org/coha/`

	Uniform			Type-Uniform			Type-Exclusivity		
Class	WN	CN	DBpedia	WN	CN	DBpedia	WN	CN	DBpedia
Authority / Subversion	.60	.76	.45	.58	.76	.48	.58	.75	.49
Care / Harm	.72	.93	.53	.71	.92	.54	.71	.92	.55
Fairness / Cheating	.59	.96	.55	.59	.96	.55	.59	.96	.53
Loyalty / Betrayal	.59	.90	.35	.59	.90	.33	.61	.90	.37
Purity / Degradation	.54	.87	.61	.53	.89	.63	.53	.87	.63
General morality	.71	.92	.68	.70	.91	.76	.68	.90	.72
Overall	.64	.88	.52	.63	.88	.53	.63	.87	.54

Table 5: Accuracy for all moral foundations, KGs and PPR methods

	Authority / Subversion	Care / Harm	Fairness / Cheating	Loyalty / Betrayal	Purity / Degradation	General morality
Authority / Subversion	.60*	.09	.05	**.11**	.07	.07
Care / Harm	.05	.72*	.04	**.07**	.05	.06
Fairness / Cheating	.06	.06	.59*	.09	.06	**.14**
Loyalty / Betrayal	**.13**	.11	.10	.59*	.03	.03
Purity / Degradation	.07	.11	.04	.00	.54*	**.24**
General morality	.04	.05	.05	.01	**.15**	.71*

(a) KG: WordNet; Method: Uniform PPR

	Authority / Subversion	Care / Harm	Fairness / Cheating	Loyalty / Betrayal	Purity / Degradation	General morality
Authority / Subversion	.49*	.08	**.13**	.10	.08	.11
Care / Harm	.10	.55*	**.11**	.10	.10	.04
Fairness / Cheating	.08	.03	.53*	**.17**	.03	.15
Loyalty / Betrayal	.14	.14	**.16**	.37*	.04	.14
Purity / Degradation	.07	.02	.05	.05	.63*	**.19**
General morality	.00	.00	**.20**	.00	.08	.72*

(b) KG: DBpedia; Method: Type-Exclusivity PPR

Table 6: Confusion matrices for predicting the moral foundation

ConceptNet prediction of the Fairness/Cheating dimension stands out through its very high scores. The prediction of the Loyalty/Betrayal foundation stands out for its poor scores on DBpedia. On the other side, the Purity/Degradation foundation is the foundation predicted best on DBpedia, achieving a score even higher than WordNet's prediction. Also the general morality class is captured quite well by all KGs, including on DBpedia, where its prediction is more accurate than on WordNet.

In Table 6, we present the confusion matrix of predicting the foundations for DBpedia and WordNet. As previously, since ConceptNet has very high scores, the confusion matrix is not conclusive, so we do not report it. Interestingly, in DBpedia, only Purity/Degradation is not mostly confused with Fairness/Cheating.

4.3 Results of Moral Polarity Prediction

Table 7 shows the results obtained on the moral polarity level, for each PPR version, per class as well as overall, for each of the three KGs. For this prediction, the random baseline would achieve, on average, a .5 score. Again, all methods achieve for all KGs significantly higher scores than the random baseline. On DBpedia, the prediction of virtues achieves slightly higher scores than the prediction of vices, while on WordNet and ConceptNet, the opposite holds. In ConceptNet, the prediction of vices vs. virtues achieves very high scores of .96 and .97, respectively. For comparison, Xie et al. (2019) report .93 accuracy on predicting the polarity when using Google N-grams embeddings and a 5-NN model. With COHA embeddings, the highest accuracy is .80, obtained with the Centroid model.

4.4 Qualitative Analysis

To also give an intuition of how our approach scores concepts that are not seeds, we also report the top 10 highest scored concepts that are not seeds, for the fairness/cheating and the authority/subversion foundations in Table 8.

WordNet concepts found for fairness/cheating are quite reasonable, while for subversion, there are some false positives (*jurisprudence, loyalty*). In ConceptNet, we overall find similar concepts for fairness/cheating, while some are on the wrong side, such as *bias, prejudice, fair, equal, judge*. In the authority/subversion foundation, often, the same concepts are scored high for both the vice

	Uniform			Type-Uniform			Type-Exclusivity		
Class	WN	CN	DBpedia	WN	CN	DBpedia	WN	CN	DBpedia
Virtue	.84	.96	.74	.81	.96	.74	.82	.96	.75
Vice	.86	.97	.72	.85	.97	.73	.83	.97	.73
Overall	.85	.96	.73	.83	.97	.74	.83	.96	.74

Table 7: Accuracy for moral polarity, KGs and PPR methods

F-virtue	F-vice	A-virtue	A-vice
equitable, just	subjective	servile	resistance
nonpartisan, nonpartizan	unprincipled	position, situation	unorthodox
democratic	partiality, partisanship	attitude, mental attitude	jurisprudence, law
mutual, reciprocal	disposition, inclination	follower	intractability, intractableness
disposition, inclination	act upon, influence	admirer, champion	dissent, resist
broad-minded	omission	reputable	loyalty, trueness
ism, philosophical system	corrupt	honorable, honourable	bad hat, mischief-maker
conformance, conformity	intolerant	tenderness, warmness	uncontrolled
true, truthful	ideology, political orientation	pious	provocative
just	advantage, vantage	courteous	disloyal, unpatriotic

(a) WN: For space reasons, we only show the two first words of each synset.

F-virtue	F-vice	A-virtue	A-vice
justify	judgment	ranke	heresy
bias	nonstandard	slang	ick
fair minded	inequality	person	legal
fair mindedly	judge	us	disagreement
prejudice	raptophilia	computing	outlaw
fair mindedness	out of proportion	detraditionalize	defier
just	fair	historical	law
philosophy	nonsegregational	maternal	person
right	separate	honourable	obedience
nonjustificational	equal	honorarium	us

(b) CN

F-virtue	F-vice	A-virtue	A-vice
National debt of the United States	Frank Stanford	Obedience to Authority: An Experimental View	Jerome Brailey
Freedom of Religion	Persecution of Ahmadis	What Comes After Goodbye	Petty treason
Work motivation	Princelings	Suprematism	Siege of Lier (1582)
Coretta Scott King	Racial wage gap in the United States	Standings	Private Lies (book)
Al-Baqara 256	Blacklisting	Robert Holden (author)	Descent
A Critique of Pure Tolerance	Shunning	Filial piety	Version Control Example
Zechariah Chafee	United States	Blondes (John Stewart album)	Sedation
Horizontal inequality	Mick Moore (political economist)	Emil Hassler	Civil Rights Act of 1968
Life estate	Barbara Risman	United States	Universum (band)
Human rights in the Middle East	Lahore Grammar School Multan	Legal Legitimacy	The Politics of Religious Apostasy

(c) DBpedia

Table 8: Top-10 non-seed concepts for every KB

and virtue (*us*), while some that are rather on the wrong side (*detraditionalize, obedience, law*).

DBpedia ranks high entities that are much different from those of WordNet and ConceptNet. While the high scores of some entities are not easily understandable (*i.e. National debt of the United States, Frank Stanford*), others nicely capture some background knowledge about the entities: i.e., *Coretta Scott King* and *Zechariah Chafee* both espoused civil rights. Interestingly, just as for ConceptNet, the *United States* is scored as highly relevant for multiple traits. This is because many DBpedia concepts that we linked to the MFD through the disambiguation pages are related to the United States, for instance, *segregation* (Fairness-vice) has been linked among others to *Housing_segregation_in_the_United_States* and *Residential_segregation_in_the_United_States*.

5 Conclusion

In this work, we investigated how moral traits, foundations, and polarity based on the MFT are captured in three widely used KGs. Our analysis reveals big differences between the three explored KGs, both quantitatively and qualitatively. ConceptNet achieves high accuracies at predicting the class of seeds in a leave-one-out evaluation. The results on WordNet are well aligned with results obtained by related work in a similar evaluation. Lastly, the seed class prediction accuracy of DBpedia scores last among the three datasets, but still significantly higher than random, and it comes with the advantage of dealing with entities such as people and organizations.

All KGs manage to accurately discriminate between virtues and vices, which is already a great step towards automatically *telling the good from the bad*. The more complex problem of predicting the foundation and the granular trait can still undergo substantial improvements, particularly on WordNet and DBpedia. However, given that our method is completely unsupervised, using just Personalized PageRank, we conclude that there is great potential in bringing morality common-sense into knowledge graphs. As future work, we are committed to further analyzing our approach, particularly on applications such as classification of texts with respect to moral values.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project ExpLAIN, Grant Number STU 266/14-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999), as well as by the Natural Sciences and Engineering Research Council of Canada.

References

Avnika B Amin, Robert A Bednarczyk, Cara E Ray, Kala J Melchiori, Jesse Graham, Jeffrey R Huntsinger, and Saad B Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Nicholas Bowman, Robert Joel Lewis, and Ron Tamborini. 2014. The morality of may 2, 2011: A content analysis of us headlines regarding the death of osama bin laden. *Mass Communication and Society*, 17(5):639–664.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar. Association for Computational Linguistics.

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

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

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

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

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. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

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.

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

T. H. Haveliwala. 2003. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796.

Ioana Hulpuş, Narumol Prangnawarat, and Conor Hayes. 2015. Path-based semantic relatedness on

linked data and its use to word and entity disambiguation. In *The Semantic Web - ISWC 2015*, pages 442–457, Cham. Springer International Publishing.

Rishemjit Kaur and Kazutoshi Sasahara. 2016. Quantifying moral foundations from various topics on twitter conversations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2505–2512.

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

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pegah Nokhiz and Fengjun Li. 2017. Understanding rating behavior based on moral foundations: The case of yelp reviews. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3938–3945.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA. Association for Computational Linguistics.

Eyal Sagi and Morteza Dehghani. 2014. Moral rhetoric in twitter: A case study of the u.s. federal shutdown of 2013. *Cognitive Science*, 36.

Livia Teernstra, Peter van der Putten, Liesbeth Noordegraaf-Eelens, and Fons Verbeek. 2016. The morality machine: Tracking moral values in tweets. In *Advances in Intelligent Data Analysis XV*, pages 26–37, Cham. Springer International Publishing.

Christopher Wolsko, Hector Ariceaga, and Jesse Seiden. 2016. Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65:7–19.

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