Uncovering Values: Detecting Latent Moral Content from Natural Language with Explainable and Non-Trained Methods

Luigi Asprino, Stefano De Giorgis and **Aldo Gangemi**[*]
Università degli Studi di Bologna, Italy
name.surname@unibo.it

Luana Bulla, Ludovica Marinucci and **Misael Mongiovì**
ISTC - Consiglio Nazionale delle Ricerche, Rome and Catania, Italy
name.surname@istc.cnr.it

Abstract

Moral values as commonsense norms shape our everyday individual and community behavior. The possibility to extract moral attitude rapidly from natural language is an appealing perspective that would enable a deeper understanding of social interaction dynamics and the individual cognitive and behavioral dimension. In this work we focus on detecting moral content from natural language and we test our methods on a corpus of tweets previously labeled as containing moral values or violations, according to Moral Foundation Theory. We develop and compare two different approaches: (i) a frame-based symbolic value detector based on knowledge graphs and (ii) a zero-shot machine learning model fine-tuned on a task of Natural Language Inference (NLI) and a task of emotion detection. Our approaches achieve considerable performances without the need for prior training.

1 Introduction

Morality as a set of social and acceptable behavioral norms (Haidt, 2012) is part of the commonsense knowledge that determines dynamics of action among social agents in areas like societal interaction (Haidt, 2001), individual conception of rightness and wrongness (Young and Saxe, 2011), moral taste and emotions (Graham et al., 2009), political commitment (Clifford and Jerit, 2013), public figure credibility (Graham et al., 2012) and narratives for explainable causal dependence of events or processes (Forbes et al., 2020).

Understanding this pervasive moral layer in both in person and *onlife* (Floridi, 2015) interaction occurrences constitutes a pillar for a good integration of AI systems in human societal communication and cultural environment. However, the difficulties in identifying data with a latent moral content, as well as cultural dependence, political orientation and the inherent subjectivity of the annotation work, make this an especially tough undertaking. In our work we aim at addressing these critical issues in the most versatile and transparent way and, to the best of our knowledge, the two approaches we propose are unprecedented in moral values detection.

The first approach employs a zero-shot learning technique. This concerns a problem setup in which a model performs classification on labels it has never seen before. By correctly interpreting the meaning of the labels and text, the classifier decides the truth value of any incoming label. This opens to the fulfillment of tasks with controversial or scarce data. We enhance the model by adding to the original text some meaningful information concerning the emotional component.

The second approach is based on an unsupervised and domain-independent system which leverages semantic web technologies and existing linguistic resources. The implementation of this method meets the suggested explainability criteria by providing a semantic knowledge graph capable of clearly describing both lexical and conceptual triggers behind the prediction. Finally we test both methods on a relevant Twitter dataset previously labeled with Graham and Haidt's Moral Foundation Theory (MFT) (Graham et al., 2013).

Our key contributions are as follows:

- We evaluate a Zero-shot learning technique based on Natural Language Inference to detect latent moral values in unstructured linguistic data.

[*] The authors are listed in alphabetical order.

Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 33 - 41
May 27, 2022 ©2022 Association for Computational Linguistics

- We enhance the zero-shot technique by the addition of the emotional component detected in the input text. We further improve the results by combining the two methods (with and without emotions).

- As an alternative method, we propose a frame-based approach based on an unsupervised and domain-independent system that guarantee explainability in reading the results achieved.

- We evaluate the above approaches on a benchmark dataset for moral values based on Twitter data and discuss the results.

The paper is organized as follows. Section 2 summarizes the results achieved in this field at the current state-of-the-art. In Section 3 we describe some baseline models, tools and resources used in our methods. Section 4.1 briefly describes the Moral Foundation Theory theoretical background, while Section 4.2 and 4.3 focus on the Zero-shot and the frame based methods, respectively. In Section 6 results of the evaluation on a manually annotated Twitter dataset are provided and discussed, while in Section 7 we delineate some possible future improvements.

2 Related Works

Previous work on identifying moral values of MFT in texts was based on word count (Fulgoni et al., 2016) or used features based on embodiments of words and sequences (Garten et al., 2016; Kennedy et al., 2021). More generally, we have observed that the most common methodological approaches in this field are divided into unsupervised and supervised methods. Unsupervised methods rely on systems not supported by external framing annotations. This approach includes architectures based on the Frame Axis technique (Kwak et al., 2021), such as those of Mokhberian and colleagues (Mokhberian et al., 2020) and Priniski and colleagues (Priniski et al., 2021). This type of approach projects words onto microframe dimensions characterized by two opposing sets of words. A framing score Moral Foundations captures the ideological and moral inclination of the texts examined. Part of the studies take as a point of reference the extended version of the Moral Foundation Dictionary (MFD) (Hopp et al., 2021), which consists of words concerning the virtues and vices of the five dyads of MFT and a sixth dimension relating to the terms of general morality. The contribution of Kobbe and colleagues (Kobbe et al., 2020), which aims to link MFD entries to WordNet in order to extend and disambiguate the lexicon, is also placed in a dictionary-based approach framework. Another unsupervised approach is explained by the work of Hulpus and colleagues (Hulpuș et al., 2020), who provide a way to explore how moral values are captured by Knowledge Graphs. The study investigates and evaluates the relevance of the entities contained in WordNet 3.1, ConceptNet and DBpedia with respect to the MFT.

Supervised methods aim to create and optimize frameworks based on external knowledge databases. The main datasets in this field are: (i) the textual corpus (Johnson and Goldwasser, 2018), which contains 93,000 tweets from US politicians in the years 2016 and 2017, and (ii) the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020), which consists of 35,000 Tweets from 7 distinct domains. In this context, the work of Roy and colleagues (Roy and Goldwasser, 2021) extends the dataset created by Johnson and Goldwasser (Johnson and Goldwasser, 2018) and applies a methodology for identifying moral values based on DRaiL, a declarative framework for deep structured prediction proposed by Pacheco and Goldwasser (Pacheco and Goldwasser, 2021). The approach adopted is mainly based on the text and information available with the unlabeled corpus such as topics, political affiliations of the authors and time of the tweets.

Our research focuses on the use of unsupervised methods. In particular, our frame-based approach is close to the work of Hulpus and colleagues (Hulpuș et al., 2020) for the use of knowledge graphs to explore latent moral (and semantic) content. However, our work enables a greater degree of knowledge integration due to disambiguation of lexical units, frame evocation, factual knowledge integration and foundational alignments, part of the text exploration process through the creation of a knowledge graph. Finally, our work provides an alternative to Frame Axis's technique (Kwak et al., 2021). Nevertheless, unlike this methodology, which implements a method based on a predefined set of terms suited for the task, we use a technology that has no a priori affinity with the suggested work. This allows us to overcome the drawbacks of utilizing a well-defined dictionary as the foundation for the entire approach and investigate the more advanced possibilities offered by an unsupervised method.

3 Reference Models

We employ a Zero-shot model based on the method developed by (Yin et al., 2019), which involves the use of pre-trained NLI models as ready-made zero-shot sequence classifiers. The approach works by using the input text as an NLI premise to classify the sequence and by developing a hypothesis starting from every possible label. In particular, the authors discuss three different aspects of classification: topic, emotion and situation detection. For each task, the model is subjected to two distinct principles: (i) *Label-partially-unseen*, where labels concerned are partially exposed to the model during a further training step, and (ii) *Label-fully-unseen*, in which the model is completely unaware of the categories. Given the lack of a specific training phase, the second approach is particularly useful in the absence of large amounts of good quality data that can be used during model implementation.

Our frame-based value detector model is based on knowledge graph generation from natural language using the FRED tool (Gangemi et al., 2017) enriched with knowledge from Framester (Gangemi et al., 2016) as a strongly connected RDF/OWL (Motik et al., 2009) knowledge graph that can be queried via its online SPARQL endpoint[1]. FRED (Gangemi et al., 2017) is a system for hybrid knowledge extraction from natural language, based on both statistical and rule-based components, which generates RDF/OWL knowledge graphs, embedding entity linking, word-sense disambiguation, and frame/semantic role detection.

Framester is a linked data hub that provides a formal semantics for frames (Gangemi, 2020), based on Fillmore's frame semantics (Fillmore, 1982). It creates/reengineers linked data versions of linguistic resources, such as WordNet (Miller, 1995), OntoWordNet (Gangemi et al., 2003b), VerbNet (Schuler, 2005), BabelNet (Navigli and Ponzetto, 2010), etc, jointly with factual knowledge bases (e.g. DBpedia (Auer et al., 2007), YAGO (Suchanek et al., 2007)). Framester also includes ImageSchemaNet (De Giorgis et al., 2022), a cognitive layer connecting image schematic sensorimotor patterns to the above-mentioned linguistic resources.

Recently, a novel layer, ValueNet[2], has been added on top of Framester. It includes moral and cultural values, and formalizes Haidt's (Graham et al., 2013) and Curry's theories (Curry et al., 2021), aligning values to Framester frames, along with a foundational ontology backbone, i.e. DOLCE-Zero (Gangemi et al., 2003a).

4 Methods

4.1 Theoretical Grounding

Through the reuse of ValueNet, our work solely focuses on Haidt's Moral Foundation Theory (MFT). MFT is grounded on the idea that, while morality could vary widely in its extension (for example, what is considered a harmful or caring behavior depends on geographical, temporal, cultural and many others dimensions), its intension presents some recurring patterns that allow to delineate a psychological system of "intuitive ethics" (Graham et al., 2013). MFT is "a nativist, cultural-developmentalist, intuitionist, and pluralist approach to the study of morality" (Graham et al., 2013): "nativist" in its neurophysiological grounding; "cultural-developmentalist" in including environmental variables in the morality-building process; "intuitionist" in declaring that there is no unique moral or non-moral trigger, but rather many patterns combining in a rationalized judgment; "pluralist" in considering that more than one narrative could fit the moral explanation process. At the core of MFT there are six dyads of values and violations:

- *Care / Harm*: a caring versus harming behavior, it grounds virtues of gentleness, kindness and nurturance.
- *Fairness / Cheating:* this foundation is based on social cooperation and typical nonzero-sum game theoretical situations based on reciprocal altruism. It underlies ideas of justice, rights and autonomy.
- *Loyalty / Betrayal:* this dyad is based on the positive outcome coming from cohesive coalition, and the ostracism towards traitors.
- *Authority / Subversion:* social interactions in terms of societal hierarchies, it underlies ideas

[1] http://etna.istc.cnr.it/framester2/sparql

[2] Available via querying Framester SPARQL endpoint: http://etna.istc.cnr.it/framester2/sparql or here: https://github.com/StenDoipanni/ValueNet

of leadership and deference to authority, as well as respect for tradition.

- *Purity / Degradation:* derived from psychology of disgust, it implies the idea of a more elevated spiritual life, it is expressed via metaphors like "the body as a temple", including the more spiritual side of religious beliefs.

- *Liberty / Oppression:* it expresses the desire of freedom and the feeling of oppression when it is negated.

4.2 Zero-shot Models

Starting from the method developed by Yin et al. (2019), we adapt a checkpoint for BART-large[3] trained on the MultiNLI (MNLI) dataset (Kim et al., 2018). Since this model has been shown to perform well for topic labeling (Khan and Chua, December (2021) and for claim verification (Reddy et al., 2021), it is a reasonable candidate for our task.

In the first step, we examine the input text for any concept similarities between its content and the moral values denoted by the labels. To the premise represented by the original textual data, we place side by side the categories suggested by Haidt's taxonomy as plausible hypotheses. In other words, we verify how much every value in the MFT's set is semantically related to every tweet in the test set (e.g. we evaluate if the concept "care" is expressed in the text "Commitment to peace, healing and loving neighbors. Give us strength and patience."). The same tweet is flanked by all the remaining moral values in the same way. The structure is based on the technique of using pre-trained NLI models as ready-made zero-shot sequence classifiers to develop a hypothesis from every possible label. As the output of the classification, results are acquired according to the predicted degree of entailment. The result of the categorization is represented by labels with a compliance score of 90% or above.

In the second step, we improve the model's prediction performances by adding more information on the latent emotional component in the original text. The input premise was subjected to an emotional detection by a model trained for this purpose[4] and then augmented by the identification of the valence of the attitude represented. For example, given the tweet "Peace, Love And Unity <3" represented as a premise, we add to this text both (i) an emotion perception component such as "This sentence is about joy sentiment." and (ii) an information about its polarity "This is positive.".

In the third step, we combine the first and second methods by unifying the prediction results to increase the likelihood of success in the classification task. In this case the results achieved by the first step and the second step were compared, assuming as the final output of the classification the moral values envisaged by either approaches (i.e. the tweet "Prayers to our brave DPD officers! We support you!" was labeled "care" and "loyalty" during the first step and only "care" during the second. In this case, the third method takes as output both labels provided, hence "care, loyalty").

All these strategies assume that artificial intelligence models can capture the interactions and connections of social groups, as well as information about individuals. Consequently, it is argued that a model might be able to draw a line of similarity between morally connoted words and ideas depending on the lexical information provided in the training phase not directly attributable to a classification method.

4.3 Frame-based Value Reasoner

The frame-based value reasoner is a tool based on a frame semantics approach (Fillmore, 1982). Its pipeline consists of the following three main steps. The first one is knowledge graph generation from natural language: the input sentence is passed to FRED, which returns a knowledge graph that includes detected FrameNet frames and frame elements, VerbNet roles, and linking to DBpedia entities and WordNet synsets.

The second step consists in the actual moral value detection: relevant entities from FRED's knowledge graph are used to query Framester SPARQL endpoint in order to link the entities extracted by FRED to MFT moral values. The full graph and an extended description of the Moral Value ontological module in Framester are available on the ValueNet github repository.[5] The resulting knowledge graph is an enrichment of the original FRED graph with MFT moral values. If

[3] https://huggingface.co/facebook/bart-large-mnli

[4] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

[5] ValueNet is available via Framester SPARQL endpoint: http://etna.istc.cnr.it/framester2/sparql and here: https://github.com/StenDoipanni/ValueNet

no value or violation is detected, the sentence is labeled as “non-moral”.

This value detection process is heuristically transparent, since it keeps track of triggering elements (e.g. synset, linked entity, frame evocation, lexical unit, etc.), so providing a fully explainable moral value detector.

5 Experiments and Results

To examine the effectiveness of our approaches in the moral value detection task, we focus on the challenge of recognizing them in the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020).

The dataset, consisting of 35k tweets, is organized into seven distinct thematic topics covering a wide range of moral concerns. Each tweet is labeled from three to eight different annotators trained to detect and categorize texts following the guidelines outlined by Moral Foundation Theory. The MFTC includes ten different moral value categories, as well as a label for textual material that does not evoke a morally meaningful response. To account for their semantic independence, each tweet in the corpus was annotated with both values and violations. To set performance baselines, we treat the annotations of the tweets by calculating the majority vote for each moral value, where the majority is considered 50% (i.e. tweet "I have no respect for the *home run king*" is labeled by four different annotators. Two of them regard the text as "non-moral" while the others as "subversion". Hence, we consider the tweet labeled as "non-moral, subversion" because each of these labels corresponds to 50% of the annotation).

Table 1 shows the results obtained by our tools on a subset of 6,075 items representing the MFTC test set. We did not include the rest of the corpus in the evaluation since the process is time consuming, considering that the code is not optimized for efficiency. Each tool is evaluated in terms of precision, recall and F1 score in predicting each label. The overall results (All in the bottom) are calculated by averaging over all labels weighted by the support (i.e. the number of elements in the ground truth with each specific label). The choice to perform the tests on a small sample of the total dataset depends on the high data processing times of the FRED-based method and the ongoing goal of a comparison with a supervised approach. This methodology would require the use of a large part of the data contained in the MFTC during the model training phase.

The presented tests are carried out by evaluating different combinations suggested by the models mentioned in Sect. 4). In particular, the Emotion-Zero-shot model displays the results obtained by exposing the Zero-shot model to an input text that has had its emotional component explained. The Emotion-Zero-shot+ architecture refers to the combination of the two methods mentioned above and corresponds to the third approach discussed in Sect. 4.2. The frame-based system recalls the results obtained from the application of the tool described in Sect. 4.3.

Given the lack of a reasonable state-of-the-art baseline of non-trained systems, we report a Random lower-bound, obtained by predicting each label with a probability corresponding to the fraction of entries in the ground truth represented by the test set with that label. Finally, in Table 1 there is no reference to the *Liberty / Oppression* dyad. This happens coherently to the lack of this label in the MFTC, due to the late introduction of this value / violation opposition in an updated version of the MFT. Triggers of this dyad are still detected by the frame-based model, and could be explored in the extended file [6], since the Liberty and Oppression knowledge graphs are part of ValueNet, but they are not considered in the evaluation metrics.

Furthermore, since the original dataset is annotated considering a 50% percentage of agreement among annotators, some of the sentences shows a combination composed by “non-moral” + some other value or violation. While for the Zero-shot models the “non-moral” label is used as a feature itself, the combination of non-morality and any kind of morality was in conflict with the conceptual structure of the frame-based detector. We therefore modified the original dataset eliminating the “non-moral” label while co-occurring with some value or violation, and repeated the experiment. The results of all the applied methods can be explored in their extended files[7].

Although performances differ, the two methods perform similarly in terms of F1, with an overall score of 45%. Specifically, Emotion-Zero-shot+ and Frame-based outperform the other models for four out of eleven labels, with F1 scores ranging from 0.12 to 0.53 for the first and from 0.11 to 0.50

[6] https://github.com/StenDoipanni/MoralDilemmas

[7] https://github.com/StenDoipanni/MoralDilemmas

Moral Value	Metric	Random	Zero-shot	Emotion-Zero-shot	Emotion-Zero-shot+	Frame-based
Care	Precision	.09	.29	.51	.29	.29
	Recall	.18	.63	.36	.69	.57
	F1-score	.11	.40	**.42**	.41	.39
Harm	Precision	.13	.30	.31	.29	.39
	Recall	.24	.80	.59	.82	.70
	F1-score	.17	.44	.41	.43	**.50**
Purity	Precision	.04	.07	.10	.07	.18
	Recall	.08	.28	.30	.32	.20
	F1-score	.05	.11	.15	.12	**.19**
Degradation	Precision	.04	.12	.15	.12	.45
	Recall	.09	.63	.30	.66	.11
	F1-score	.06	**.20**	**.20**	**.20**	.18
Loyalty	Precision	.07	.40	.73	.40	.40
	Recall	.15	.45	.14	.46	.30
	F1-score	.10	.42	.24	**.43**	.34
Betrayal	Precision	.05	.17	.37	.17	.57
	Recall	.10	.44	.29	.44	.17
	F1-score	.07	.25	**.32**	.25	.27
Fairness	Precision	.07	.60	.85	.58	.16
	Recall	.15	.47	.26	.48	.11
	F1-score	.09	**.53**	.40	**.53**	.13
Cheating	Precision	.11	.54	.64	.54	.75
	Recall	.22	.29	.19	.29	.28
	F1-score	.15	.38	.30	.38	**.41**
Authority	Precision	.04	.17	.40	.18	.15
	Recall	.08	.28	.04	.29	.08
	F1-score	.05	.21	.07	**.22**	.11
Subversion	Precision	.08	.20	.15	.17	.28
	Recall	.16	.36	.39	.40	.17
	F1-score	.11	**.25**	.21	.24	.21
Non-moral	Precision	.44	.40	.46	.47	.59
	Recall	.66	.28	.86	.91	.72
	F1-score	.53	.33	.60	.62	**.65**
All	Precision	.22	.35	.46	.38	.47
	Recall	.36	.41	.52	.67	.48
	F1-score	.27	.35	.42	**.45**	.44

Table 1. Precision, Recall and F1 score for each model on the MFTC dataset.

for the second. These two architectures result in an improvement of 10 % compared to the Emotion-zero-shot model and 20 % compared to the Zero-shot model, and they performs vastly better than Random.

6 Discussion

As expected, the results for the single labels vary according to the difficulties encountered by classifiers in the interpretation of their meaning. For example, moral values such as "Harm" or "Care" convey more generic content and are therefore easier to identify. Conversely, concepts like "Degradation" or "Subversion" contain shades of meaning that are more difficult to grasp.

The results drawn from the Zero-shot models make this problem evident and difficult to solve as the intrinsic nature of machine learning models does not encompass a direct understanding of their decision-making phases. One possible solution would be to subject the models to few-shot learning, which is a fine-tuning with a little amount of data relevant for the moral values detection task. However, this would not be part of our main need, which is to develop flexible approaches that do not require training. Despite the task's complexity, the results imply that not only can moral values be detected in natural language texts, but also that models developed for NLI may be adapted to other tasks through the unintentional acquisition of abstract conceptions and concepts connected to the field of social value.

Results obtained from the frame-based value detector are provided as additional material[8]. Value triggers are listed in the "trigger" column, while value detection is shown in the "prediction" column. The full knowledge graph can be retrieved by passing the tweet content in column "tweet_text" as input to the FRED online demo[9], ticking the "align to Framester" option.

A necessary caveat is that, being the value labeling a subjective task, a certain amount of disagreement should always be taken into account. In this regard, the detection shows better results on those values whose extension seems more generic, e.g. a more broad concept like "harm", than a more opaque one like "purity", as described in Sect. 4.1. Additionally, the performance results could depend on two factors. The first factor is the success of the FRED tool in producing a knowledge graph from a fragmented syntax like the one used in tweets. In fact, even when a well formed graph is produced, if the value trigger is not in the main sentence e.g. it is an adjective of a pronoun in a subordinate sentence, it is possible that its disambiguation / frame evocation is not shown in the graph, due to internal FRED saliency heuristics. The second factor is that human value labeling is a task carried out with a certain subjective threshold. If we consider the example: "Horrible amount of anti-Islam bigotry are Paris attacks. ISIS murder more MUSLIMS than anyone else.", value labels for this sentence are "cheating" and "harm", while the detector predicts "cheating", "harm" and "purity". This happens because, along with triggers like the `fs:Offenses` and `fs:Killing` Framester frames, `wn:murder-noun-1` WordNet synset and the `dbr:Bigotry` DBpedia entity, the DBpedia entry `dbr:Muslim` is also retrieved, which according to "purity" definition (see Sect. 4.1) covers the semantics of a more spiritual aspect of life, and it is therefore a "purity" trigger.

7 Conclusions and Future Work

In our work we detect latent moral content from natural language in a versatile and transparent way, proposing two approaches (zero-shot and heuristic) that do not require training. The approaches assume Haidt's Moral Foundation Theory as a reference for moral values, and have been tested on the Moral Foundation Twitter Corpus.

Results are unprecedented in using domain independent methods. Future work will include improving the performance of the Zero-shot models through the creation of a technique capable of comprehending the intricacies of the most contentious moral values. Furthermore, we plan to build an implementation that gives greater weight to the most significant aspects of the sentence, in order to more simply detect the prevailing moral value.

For the frame-based value detector more precise results could be achieved via different refinements such as a set of heuristics based on the syntax, and consequently on the frame structure of the sentence, which would allow new and more complex inferences. The commitment to some value could, for example, be expressed by the negation of the value violation, or via a negative polarity of a verb which takes as argument some value trigger. Some

[8] https://github.com/StenDoipanni/MoralDilemmas

[9] http://wit.istc.cnr.it/stlab-tools/fred/demo/

other possibility to improve the results could be in a quantitative or qualitative way, namely introducing a scoring system based on the amount of trigger occurrences per value, or weighting differently the type of trigger (WordNet synset, FrameNet frame, etc.).

Finally, an interesting possibility is to conjugate the approaches and this could drive to various possibilities, for example the introduction of a layer in the aforementioned value trigger scoring system, able to guide the prediction of the final output, as well as using the knowledge base, in particular the extended lexical coverage from ValueNet graphs to improve Zero-shot models performance. A possible way could be to analyze the frame responsible for the value triggering by measuring its relevance inside the sentence via machine learning techniques. To conclude, further experiments can be done on different types of datasets as well as extending the employed dataset, and to compare with different methodological approaches, including supervised methods.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

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.

Oliver Scott Curry, Mark Alfano, Mark J Brandt, and Christine Pelican. 2021. Moral molecules: Morality as a combinatorial system. *Review of Philosophy and Psychology*, pages 1–20.

Stefano De Giorgis, Aldo Gangemi, and Dagmar Gromann. 2022. Imageschemanet: Formalizing embodied commonsense knowledge providing an image-schematic layer to framester. *Semantic Web Journal*, forthcoming.

Charles J Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–138. Seoul: Hanshin.

Luciano Floridi. 2015. *The onlife manifesto: Being human in a hyperconnected era*. Springer Nature.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736.

Aldo Gangemi. 2020. Closing the loop between knowledge patterns in cognition and the semantic web. *Semantic Web*, 11(1):139–151.

Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: a wide coverage linguistic linked data hub. In *European Knowledge Acquisition Workshop*, pages 239–254. Springer.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003a. Sweetening wordnet with dolce. *AI magazine*, 24(3):13–13.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003b. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 820–838. Springer.

Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893.

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.

Jesse Graham, Brian A Nosek, and Jonathan Haidt. 2012. The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PloS one*, 7(12):e50092.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53(1):232–246.

Ioana Hulpuș, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. 2020. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. Moral concerns are differentially observable in language. *Cognition*, 212:104696.

Qaisar Khan and Huina Chua. December (2021) 46 - 59. An automated topics labeling framework using zero-shot text classification. *Journal of Engineering Science and Technology Special Issue on ACSAT*.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. Association for Computational Linguistics, ACL.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. Frameaxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics*, pages 206–219. Springer.

Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, Carsten Lutz, et al. 2009. Owl 2 web ontology language profiles. *W3C recommendation*, 27(61).

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

J Hunter Priniski, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu, and P Jeffrey Brantingham. 2021. Mapping moral valence of tweets following the killing of george floyd. *arXiv preprint arXiv:2104.09578*.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of us politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Liane Young and Rebecca Saxe. 2011. When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, 120(2):202–214.