Detection of Morality in Tweets Based on the Moral Foundation Theory

Luana Bulla[1,2(✉)], Stefano De Giorgis[3], Aldo Gangemi[1,2,3], Ludovica Marinucci[1], and Misael Mongiovì[1,2]

[1] ISTC - Consiglio Nazionale delle Ricerche, Rome, Italy
{luana.bulla,aldo.gangemi,ludovica.marinucci,misael.mongiovi}@istc.cnr.it
[2] ISTC - Consiglio Nazionale delle Ricerche, Catania, Italy
[3] Università degli Studi di Bologna, Bologns, Italy
stefano.giorgis@unibo.it

Abstract. Moral Foundations Theory is a socio-cognitive psychological theory that constitutes a general framework aimed at explaining the origin and evolution of human moral reasoning. Due to its dyadic structure of values and their violations, it can be used as a theoretical background for discerning moral values from natural language text as it captures a user's perspective on a specific topic. In this paper, we focus on the automatic detection of moral content in sentences or short paragraphs by means of machine learning techniques. We leverage on a corpus of tweets previously labeled as containing values or violations, according to the Moral Foundations Theory. We double evaluate the result of our work: (i) we compare the results of our model with the state of the art and (ii) we assess the proposed model in detecting the moral values with their polarity. The final outcome shows both an overall improvement in detecting moral content compared to the state of the art and adequate performances in detecting moral values with their sentiment polarity.

Keywords: Text classification · Moral foundation theory · Transformers

1 Introduction

Morality can be described as a set of social and acceptable behavioral norms [13] part of our every day commonsense knowledge. It underlies many situations in which social agents are requested to participate in the dynamics of actions in domains like societal interaction [12], political ideology [10] and commitment [1], individual conception of rightness and wrongness [27], public figure credibility [11], and plausible narratives to explain causal dependence of events or processes [6]. Therefore, moral values can be seen as parameters that allow humans to assess personal and other people's actions. Understanding this pervasive moral layer in both in person and *onlife* interaction occurrences [5] constitutes a pillar

G. Nicosia et al. (Eds.): LOD 2022, LNCS 13810, pp. 1–13, 2023.
https://doi.org/10.1007/978-3-031-25599-1_1

for a good integration of AI systems in human societal communication and cultural environment. Moral values detection from natural language text passages might help us better understand the cultural currents to which they belong. However, the difficulties in identifying data with a latent moral content, as well as cultural differences, political orientation, personal interpretation and the inherent subjectivity of the annotation work, make this an especially tough undertaking.

In our work, we aim at addressing these critical issues by fine-tuning a BERT-based model [4], a well-known architecture that has achieved cutting-edge performance in a variety of NLP tasks, including classification tasks [21]. We apply the BERT-based model on the dataset developed by Hoover and colleagues [14], which contains 35,000 Tweets tagged with Graham and Haidt's Moral Foundation Theory (MFT) [9]. Based on the work of Hoover and colleagues [14], we calculate the agreement between the annotators to estimate the moral values associated with each tweet in the dataset, and then validate the classification results in two distinct ways. The first one is based on the comparison of our system with the state of the art on the presence or absence in the text of the five MFT's dyads [9]. Each of the Moral Foundations consists in the union of the moral value and its violation (e.g. "Care" *and* "Harm"). The results of the classification show a noticeable improvement. The second assessment considers the polarity of the value evoked by the evaluated text, revealing with notable precision the value or its opposition (i.e. distinguishing "Care" from "Harm"). Our approach expands the set of labels, hence making the process harder. Finally, we propose an analysis of the results that separates "moral" passages from "non-moral" ones.

Our main contributions are:

- we propose a BERT-based model for the automatic detection of latent moral values in short text passages;
- we perform an extensive comparison of the proposed model with the state of the art in the task of inferring the MFT's dyadic dimensions;
- we perform an assessment of the model in detecting moral values and their violations, thus highlighting the polarity of moral sentiment, and discuss the results.

The paper is organized as follows. In Sect. 2 we provide a brief introduction to the theoretical background we rely on, i.e. Moral Foundations Theory (MFT) [9], which is adopted to perform moral value detection by previous work we refer to in Sect. 3, thus highlighting similarities and differences with our contribution. Section 4 details the description of the Moral Foundation Twitter Corpus (MFTC) [14], and our BERT-based approach. In Sect. 5, we describe our two evaluation methodologies, provide the results concerning precision, recall and F1 score for our approach in comparison with the state of the art and present and discuss a confusion matrix; Sect. 6 discusses the above mentioned results comparing them with those described in the work of Hoover and colleagues [14]. Section 7 concludes the paper envisaging possible future developments of our approach.

2 Theoretical Framework

Our work is framed on Haidt's Moral Foundation Theory (MFT). MFT is grounded on the idea that morality could vary widely in its *extension*, namely in what is considered a harmful or caring behavior, according to cultural, geographical, temporal, societal and other factors [10], but not in its *intension*, showing recurring patterns that allow to delineate a psychological system of "intuitive ethics" [9]. MFT is also considered a "nativist, cultural-developmentalist, intuitionist, and pluralist approach to the study of morality" [9]. "Nativist" due to its neurophysiological grounding; "cultural-developmentalist" because it includes environmental variables in the morality-building process [11]; "intuitionist" in asserting that there is no unique moral or non-moral trigger, but rather many co-occurring patterns resulting in a rationalized judgment [12]; "pluralist" in considering that more than one narrative could fit the moral explanation process [13].

MFT is built around a core of six dyads of values and violations:

- *Care/Harm:* caring versus harming behavior, it is grounded in mammals attachment system and cognitive appraisal mechanism of dislike of pain. It grounds virtues of kindness, gentleness and nurturance.
- *Fairness/Cheating:* it 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:* it is based on tribalism tradition and the positive outcome coming from cohesive coalition, as well as the ostracism towards traitors.
- *Authority/Subversion:* social interactions in terms of societal hierarchies, it underlies ideas 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", and includes the more spiritual side of religious beliefs.
- *Liberty/Oppression:* it expresses the desire of freedom and the feeling of oppression when it is negated. This last dyad is listed here for a complete overview of the MFT [13]. However, it is not considered in the Moral Foundation Twitter Corpus (MFTC), and thus it is not employed in our classification process, as explained in Sect. 4.

MFT's dyadic structure for defining values and their violations, which coincides with a positive vs negative polarity, and is applied to an extended labeled corpus (i.e. MFTC), offers a sound theoretical framework for the latent moral content detection task we aim to perform.

3 Related Works

Previous work on detecting MFT's moral values in text have relied on word counts [7] or have employed features based on word and sequence embeddings [8,

18]. More broadly, we observe that the most common methodological approaches in this field are divided into unsupervised and supervised methods.

The non-supervised methods rely on systems not backed by external framing annotations. This approach includes architectures based on Frame Axis technique [19], such as those by Mokhberian and colleagues [23] and Priniski and colleagues [25]. This type of approach projects the words on micro-frame dimensions characterized by two sets of opposing words. A Moral Foundations framing score captures the ideological and moral inclination of the text examined. Part of the studies take as a reference point the extended version of the Moral Foundation Dictionary (eMFD) [15], which consists of words concerning virtues and vices of the five MFT's dyads and a sixth dimension relating to the terms of general morality.

The supervised methods aim at creating and optimizing frameworks based on external knowledge databases. The main datasets in this field are: (i) the textual corpus [17], which contains 93,000 tweets from US politicians in the years 2016 and 2017, and (ii) the Moral Foundation Twitter Corpus (MFTC) [14], which consists of 35,000 Tweets from 7 distinct domains. In this context, the work of Roy and colleagues [26] extends the data set created by Johnson and Goldwasser [17] and applies a methodology for identifying moral values based on DRaiL, a declarative framework for deep structured prediction proposed by Pacheco and Goldwasser [24]. The approach adopted is mainly based on the text and information available with the unlabeled corpus such as topics, authors' political affiliations and time of the tweets.

Our research focuses on the use of supervised methods. Specifically, we are close to the work of Hoover and colleagues [14] in terms of the final goal and dataset used. Therefore, due to these similarities, in Sect. 6 we compare our results with those described in [14] following the same data processing procedures. Unlike the authors' methodology, which implement a multi-task Long Short-Term Memory (LSTM) neural network to test the MFTC dataset, we employ a more recent technology based on a transformer language model called BERT [4]. This architecture is pre-trained from unlabeled data extracted from BooksCorpus with 800 million words and English Wikipedia with 2.500 million words. BERT learns contextual embeds for words in an unsupervised way as a result of the training process. After pre-training, the model can be fine-tuned with fewer resources to optimize its performance on specific tasks. This allows it to outperform the state of the art in multiple NLP tasks.

4 Methodology

In the following we detail our approach of detecting moral foundations by applying a BERT-based model to the Moral Foundation Twitter Corpus (MFTC), which has been annotated according to the moral dyads of Haidt's MFT (cf. Sect. 2).

4.1 The Moral Foundation Twitter Corpus

The Moral Foundation Twitter Corpus (MFTC), developed by Hoover and colleagues [14], and consisting of 35k tweets, is organized into seven distinct thematic topics covering a wide range of moral concerns, relevant to current social science problems. In summary, the seven topics are:

- *All Lives Matter* (ALM), which aggregates all tweets using the hashtags #BlueLivesMatter and #AllLivesMatter published between 2015 and 2016. These materials refer to the American social movement that arose in response to the African-American community's Black Lives Matter movement.
- *Baltimore Protests* (Baltimore), which collects tweets from cities where protests over the murder of Freddie Gray took place during the 2015 Baltimore protests.
- *Black Lives Matter* (BLM), which groups all tweets using the hashtags #BLM and #BlackLivesMatter posted between 2015–2016. These data refer to the African-American community's movement born in reaction to the murders of Black people by the police forces and against discriminatory policies against the Black community.
- *2016 Presidential Election* (Election), which are scraped tweets from the followers of @HillaryClinton, @realDonaldTrump, @NYTimes, @washingtonpost and @WSJ during the 2016 Presidential election season.
- *Davidson*, whose tweets are taken from the corpus of hate speech and offensive language by Davidson and colleagues [3].
- *Hurricane Sandy* (Sandy), which includes all tweets posted before, during, and immediately after Hurricane Sandy (10/16/2012-11/05/2012).
- *#MeToo*, whose tweets contain data from 200 individuals involved in the #MeToo movement, a social movement against sexual abuse and sexual harassment.

Unlike previous datasets in this field (cf. [17,26] described in Sect. 3), the MFTC includes both issues with no connection to politics (i.e. Hurricane Sandy) and topics with political meaning (i.e. the Presidential election). Furthermore, the latter have no ideological inclinations because the facts pertain to issues that both liberals (i.e. BLM) and democrats (i.e. ALM) care about.

The corpus heterogeneity corresponds to our study goal, which is to recognize moral values that are not registered in a single area. Each tweet is labeled from three to eight different annotators trained to detect and categorize text passages following the guidelines outlined by Haidt's MFT [9]. In particular, the dataset 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. Furthermore, to set performance baselines, tweet annotations are processed by calculating the majority vote for each moral value, where the majority is considered 50%.

4.2 A BERT-Based Method for Detecting Moral Values

To identify moral values in short text we adopt a pre-trained linguistic model for the English language called SqueezeBERT [16]. This model is very similar to the BERT-based architecture [4], which is a self-attention network able to remove the unidirectionality constraint by using a masked language model (MLM) pre-training objective. This task masks some tokens from the input at random, with the goal of predicting the masked word's original vocabulary ID based solely on its context. Furthermore, BERT uses a next sentence prediction task that jointly pre-trains sentence-pair representations.

Bert-derived models have been utilized in a range of applications, with several of them demonstrating significant improvements in natural language task performance, including classification tasks [22]. In the latter, most BERT-derived networks generally consist of three stages: (i) the *embedding*, (ii) the *encoder*, and (iii) the *classifier*. The *embedding* translates pre-processed words into learned feature-vectors of floating-point values; the *encoder* consists of multiple blocks, each of them formed by a self-attention module followed by three fully connected layers, known as feed-forward network layers (FFN); and the *classifier* generates the network's final output. SqueezeBERT model architecture is comparable to BERT-base, except grouped convolutions replace the point-wise fully-connected layers. This change optimizes the BERT-based structure by making the SqueezeBERT model faster in executing the task.

To detect moral foundations in tweets, we fine-tuned SqueezeBERT by using labeled data from the MFTC. Each tweet went through a lemmatization and cleaning procedure before being categorized, removing references to links and Retweets (RT). Furthermore, all of the tweets were truncated to a length of 40 tokens before passing through BERT. We use a learning rate of 2e−5, batch size of 64, drop-out of 0.4, and AdamW as optimizer. To measure the distance of the model predictions from the real values, we adopted a commonly used loss function for multi-label classification tasks:

$$loss(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_i \cdot \log \frac{1}{1 + e^{-x_i}} + (1 - y_i) \cdot \log \frac{e^{-x_i}}{1 + e^{-x_i}} \right)$$

where N is the number of labels, x is the N-length output vector of the classifier and y is the N-length binary vector representing the real labels (1 for labels associated with the text, 0 in the other cases). The loss function compares the predicted vector x with the actual annotation y and calculates a score that penalizes probabilities that are distant from the expected value. In other words, the metric establish a criterion that optimizes a multi-label one-versus-all loss based on max-entropy, between the classifier output x and the target y.

To make the prediction, we first normalize the classifier output to return values between 0 and 1 by means of a sigmoid function. Then we set a 0.9 threshold and chose the labels whose predicted values are above or attained to the threshold.

5 Results and Evaluation

We perform an extensive experimental analysis to evaluate the performances of the proposed approach. We first compare the performances of our approach in detecting the five dyads of the MFT and the polarity of tweets (i.e. positive versus negative) with the state-of-the-art method from Hoover and colleagues [14]. Then, we present the results of detecting all moral values and their negation. All experiments were performed in the MFTC corpus introduced in Sect. 4.1.

5.1 Classification of Tweets Based on the MFT Dimentions

To examine the effectiveness of our approach in detecting moral dyadic dimensions, we compare our classifier with the results obtained by Hoover and colleagues [14] by means of employing a Long-short Term Memory (LSTM) model [2,20].

Tables 1 to 5 show the results obtained both by our model (i.e. BERT-Model) and the Hoover and colleagues' model (i.e. LSTM) on 10-fold cross-validation. We got the precision, recall, and F1 score for each of the moral dyads from Graham and Haidt MFT's taxonomy. For BERT-Model, each of the moral dimensions has been defined as the union of the two moral values that compose it (e.g. for the Care moral value, we considered as positive tweets the ones labeled either with Care or Harm). For the 10-fold cross-validation, each fold was obtained by splitting the set of tweets of each subtopic (i.e. ALM, #MeToo, Sandy) in 10 parts, after shuffling the tweets randomly. For LSTM we report the values from Hoover and colleagues [14].

We also evaluated the performances in detecting text with moral content in comparison with the scores reported by Hoover and colleagues [14] (cf. Table 6). We distinguished positive and negative tweets in the following way: items labeled with at least one of the ten moral values are considered as part of the positive set, while non-moral tweets were considered as part of the negative set (Table 2, 3 and 4).

With an F1 ranging from 87% to 81%, the data show a noticeable performance improvement over LSTM in all situations, with a few exceptions. For the whole dataset (column All), precision, recall and F1 are above or attained to 80% for all labels while for LSTM these values are usually between 28% and

Table 1. Classification results for the Care dimension of the MFT.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
LSTM	F1	.63	.65	.26	.77	.61	.06	.36	.78
	Precision	.81	.80	.76	.86	.78	.64	.69	.81
	Recall	.52	.55	.16	.70	.50	.03	.25	.75
BERT-Model	F1	**.82**	**.86**	**.65**	**.91**	**.81**	**.23**	**.83**	**.81**
	Precision	**.86**	**.88**	**.81**	**.92**	**.85**	**.79**	**.87**	**.86**
	Recall	**.82**	**.86**	**.63**	**.91**	**.81**	**.35**	**.82**	**.81**

Table 2. Classification results for the Fairness dimension of the MFT.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
LSTM	F1	.70	.75	.43	.88	.75	.02	.55	.10
	Precision	.81	.84	**.81**	**.91**	.85	.35	.76	.06
	Recall	.61	.68	.30	.86	.68	.01	.43	**.87**
BERT-Model	F1	**.81**	**.87**	**.60**	**.89**	**.83**	**.21**	**.77**	**.81**
	Precision	**.85**	**.89**	.74	**.91**	**.87**	**.86**	**.79**	**.87**
	Recall	**.80**	**.87**	**.60**	**.88**	**.83**	**.26**	**.77**	.79

Table 3. Classification results for the Loyalty dimension of the MFT.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#Me Too	Sandy
LSTM	F1	.70	.75	.43	.88	.75	.02	.55	.10
	Precision	.81	.84	**.81**	.91	.85	.35	.76	.06
	Recall	.61	.68	.30	.86	.68	.01	.43	**.87**
BERT-Model	F1	**.85**	**.92**	**.46**	**.95**	**.85**	**.27**	**.86**	**.79**
	Precision	**.88**	**.93**	.66	**.95**	**.89**	**.81**	**.87**	**.88**
	Recall	**.83**	**.91**	**.55**	**.95**	**.84**	**.35**	**.86**	.75

Table 4. Classification results for the Authority dimension of the MFT.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#Me Too	Sandy
LSTM	F1	.47	.57	.19	.83	.33	.01	.47	.59
	Precision	.80	.85	.77	.91	.80	.24	.67	.80
	Recall	.34	.43	.11	.76	.21	.01	.36	.46
BERT-Model	F1	**.82**	**.92**	**.59**	**.96**	**.87**	**.19**	**.68**	**.80**
	Precision	**.87**	**.93**	**.87**	**.96**	**.92**	**.89**	**.72**	**.86**
	Recall	**.80**	**.91**	**.50**	**.96**	**.85**	**.22**	**.67**	**.79**

Table 5. Classification results for the Purity dimension of the MFT.

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
LSTM	F1	.41	.57	.07	.48	.47	.04	.53	.15
	Precision	.80	.85	.81	.81	.79	.48	.71	.72
	Recall	.28	.43	.03	.34	.33	.02	.43	.09
BERT-Model	F1	**.87**	**.95**	**.89**	**.97**	**.84**	**.16**	**.71**	**.91**
	Precision	**.91**	**.97**	**.94**	**.97**	**.86**	**.83**	**.77**	**.96**
	Recall	**.86**	**.94**	**.86**	**.97**	**.83**	**.26**	**.71**	**.88**

81%. For both models, performances varied across the discourse domain. This is notably true in the sub-corpus "Davidson" that includes hate messages from Davidson and colleagues's corpus of hate speech and offensive language [3]. In

Table 6. Classification results for the Moral Sentiment

Model	Metric	All	ALM	Baltimore	BLM	Election	Davidson	#MeToo	Sandy
LSTM	F1	.80	.76	.69	**.89**	.77	.14	.81	.86
	Precision	.81	.77	.81	.86	.78	.49	.78	**.97**
	Recall	.79	.76	.61	**.92**	.76	.08	**.84**	.77
BERT-Model	F1	**.85**	**.83**	**.86**	.88	**.78**	**.89**	**.83**	**.88**
	Precision	**.87**	**.85**	**.89**	**.90**	**.80**	**.89**	.84	.90
	Recall	**.86**	**.85**	**.86**	.89	**.79**	**.88**	**.84**	**.90**

this scenario, the majority of tweets have several flaws and are labeled with a high numbers of 'non-moral' labels that amplify the misprediction outcomes.

5.2 Detection of Moral Values with Polarity

To evaluate the performance of the model in detecting both the moral values and their violations, we verify the results of the prediction by highlighting the moral sentiment polarity. This leads to an 11-class classification task (we added a "Non-moral" class to the 10 moral classes given by MFT).

Table 7 shows the results obtained by testing the model on 6,125 items of the MFTC test set. Each label is evaluated in terms of precision, recall and F1 score. The overall results (All in the bottom) are calculated by averaging over all labels weighted by their support (i.e. the number of elements in the ground truth with each specific label).

Table 7. Model F1, Precision, and Recall Scores for Moral Values classification

Moral value	Precision	Recall	F1
Care	0.76	0.73	0.75
Harm	0.67	0.69	0.68
Purity	0.63	0.52	0.57
Degradation	0.59	0.43	0.50
Non-moral	0.90	0.82	0.86
Loyalty	0.66	0.66	0.66
Betrayal	0.56	0.54	0.55
Fairness	0.83	0.72	0.77
Cheating	0.69	0.66	0.67
Authority	0.62	0.58	0.60
Subversion	0.44	0.51	0.47
All	0.76	0.71	0.73

The results reveal an F1 of 73% overall. As expected, performance varied significantly depending on the predicted moral value, with an F1 score ranging

from 47% for "subversion" to 86% for tweets classified as "non-moral". Specifically, the best results recorded for positive moral values in terms of F1 score are related to "care" (75%) and "fairness" (77%) while for negative values we have "harm" (68%) and "cheating" (67%).

Furthermore, Table 8 shows the confusion matrix generated from the above predictions. The results demonstrate how the classifier often swaps values for "non-moral". Furthermore, the greater ambiguity is given by confusing "subversion" or "betrayal" with "cheating", and "authority" with "loyalty" or "subversion".

Table 8. Confusion Matrix for predicted Moral Values

	Care	Purity	Non-moral	Loyalty	Cheating	Fairness	Subversion	Betrayal	Degradation	Harm	Authority
Care	367	12	23	34	8	7	9	1	3	26	10
Purity	14	74	29	4	0	7	0	2	3	1	2
Non-moral	43	36	2373	59	82	27	71	14	19	140	42
Loyalty	19	3	25	215	11	5	8	4	0	7	20
Cheating	2	1	46	7	411	19	50	40	10	34	4
Fairness	7	2	17	5	21	268	15	2	1	7	11
Subversion	4	0	34	7	16	0	156	32	14	24	15
Betrayal	2	0	30	10	16	0	18	70	3	11	0
Degradation	0	1	20	1	10	1	18	2	65	18	1
Harm	17	2	41	1	31	1	21	8	3	342	4
Authority	5	1	13	11	0	4	14	0	0	2	112

6 Discussion

As expected, in classifying moral values in tweets (cf. Sect. 5.2) the performance varied substantially across labels. Although the tool performed reasonably well overall, some labels appear to be interpreted inconsistently, resulting in ambiguities that relate to the message expressed in the text. This is visible where components of subversion have been mixed together with moral betrayal feelings (i.e. the tweet "Trump Isn't Hitler? Really? #DonaldTrump is another Hitler! I can't stand Dictators and Traitors! #FDT #Resist #NoH8 #EndRacism #LoveWINS" is tagged with "subversion" by annotators and classified as "betrayal" by the model).

Furthermore, a text can simultaneously communicate multiple moral values that are not identified by the annotators but recognized by the classifier (i.e. the tweet "I'm continually shocked by the stupidity of people. Support our country, support your local police, respect authority. #AllLivesMatter" is labeled only as "authority" by annotators but it is classified thought "authority" and "loyalty" by the model). The results revealed that concepts like "Degradation" or "Subversion" have shades of meaning that are more difficult to detect. This criticality, along with the annotation task's great subjectivity, led to the display activity anomalies in the human value labeling task and increase the task detection's difficulty.

The task of classifying dimensions of the Moral Foundation (cf. Sect. 5.1) turned out to be simpler and led to better results. With an overall F1 of 83%, the proposed model outperforms the state-of-the-art architecture represented by the Hoover and colleagues' work [14], which implemented and trained a multi-task Long Short-Term Memory (LSTM) neural network [2,20] to predict moral sentiment. The likelihood of misclassification is decreased in this circumstance since the number of labels to predict is substantially lower. Indeed, only 5 dyads and a value designated as non-moral are tracked, compared to the 11 labels supplied for the classification of the single moral values in text. The most common anomalies in the classification are highlighted at the sub-corpora level and sometimes worse performances of the model are related to the low cleanliness of the data.

7 Conclusion and Future Work

In our work, we detect latent moral content from natural language text by means of a BERT-based model. The approach considers Haidt's Moral Foundation Theory as a reference for moral values and employ a BERT-based classifier fine-tuned and tested on the Moral Foundation Twitter Corpus (MFTC). The results show an advancement of the state of the art presented by Hoover and colleagues [14] in detecting the dyads that compound Moral Foundations. Furthermore, we present the results of the multi-label classifier taking into account the polarity of moral sentiment and expanding the set of reference labels.

We plan to build an implementation of our model that gives greater weight to the most significant aspects of the sentence, in order to improve the detection of the prevailing moral value, especially when associated to specific emotions evocative of a certain value. Additionally, further experiments on different datasets would help to verify the performance of the model across different domains and on text with different characteristics.

Acknowledgements. We gratefully acknowledge Morteza Dehghani for supporting us in setting up the MFTC dataset and the testset for comparison. This work is supported by the H2020 projects TAILOR: Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization – EC Grant Agreement number 952215 – and SPICE: Social Cohesion, Participation and Inclusion through Cultural Engagement – EC Grant Agreement number 870811.

References

1. Clifford, S., Jerit, J.: How words do the work of politics: moral foundations theory and the debate over stem cell research. J. Politics **75**(3), 659–671 (2013)
2. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167 (2008)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)

4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2019)
5. Floridi, L.: The Onlife Manifesto: Being Human in a Hyperconnected Era. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-319-04093-6
6. Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: learning to reason about social and moral norms. arXiv preprint arXiv:2011.00620 (2020)
7. Fulgoni, D., Carpenter, J., Ungar, L., Preoţiuc-Pietro, D.: An empirical exploration of moral foundations theory in partisan news sources. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3730–3736 (2016)
8. Garten, J., Boghrati, R., Hoover, J., Johnson, K.M., Dehghani, M.: Morality between the lines: detecting moral sentiment in text. In: Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes (2016)
9. Graham, J., et al.: Moral foundations theory: the pragmatic validity of moral pluralism. In: Advances in Experimental Social Psychology, vol. 47, pp. 55–130. Elsevier (2013)
10. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. J. Pers. Social Psychol. **96**(5), 1029 (2009)
11. Graham, J., Nosek, B.A., Haidt, J.: The moral stereotypes of liberals and conservatives: exaggeration of differences across the political spectrum. PloS One **7**(12), e50092 (2012)
12. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**(4), 814 (2001)
13. Haidt, J.: The righteous mind: why good people are divided by politics and religion. Vintage (2012)
14. Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A.M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al.: Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiment. Social Psychol. Pers. Sci. **11**(8), 1057–1071 (2020)
15. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (emfd): development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behav. Res. Methods **53**(1), 232–246 (2021)
16. Iandola, F.N., Shaw, A.E., Krishna, R., Keutzer, K.W.: SqueezeBERT: what can computer vision teach nlp about efficient neural networks? arXiv:2006.11316 (2020)
17. Johnson, K., Goldwasser, D.: Classification of moral foundations in microblog political discourse. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 720–730 (2018)
18. Kennedy, B., et al.: Moral concerns are differentially observable in language. Cognition **212**, 104696 (2021)
19. Kwak, H., An, J., Jing, E., Ahn, Y.Y.: Frameaxis: characterizing microframe bias and intensity with word embedding. PeerJ Comput. Sci. **7**, e644 (2021)
20. Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114 (2015)
21. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification: a comprehensive review. ACM Comput. Surv. (CSUR) **54**(3), 1–40 (2021)
22. Mohammed, A.H., Ali, A.H.: Survey of bert (bidirectional encoder representation transformer) types. In: Journal of Physics: Conference Series, vol. 1963, p. 012173. IOP Publishing (2021)

23. Mokhberian, N., Abeliuk, A., Cummings, P., Lerman, K.: Moral framing and ideological bias of news. In: Aref, S., Bontcheva, K., Braghieri, M., Dignum, F., Giannotti, F., Grisolia, F., Pedreschi, D. (eds.) SocInfo 2020. LNCS, vol. 12467, pp. 206–219. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60975-7_16
24. Pacheco, M.L., Goldwasser, D.: Modeling content and context with deep relational learning. Trans. Assoc. Comput. Linguist. **9**, 100–119 (2021)
25. Priniski, J.H., et al.: Mapping moral valence of tweets following the killing of george floyd. arXiv preprint arXiv:2104.09578 (2021)
26. Roy, S., Goldwasser, D.: 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, pp. 1–13 (2021)
27. Young, L., Saxe, R.: When ignorance is no excuse: different roles for intent across moral domains. Cognition **120**(2), 202–214 (2011)