Recognising moral foundations in online extremist discourse
A cross-domain classification study

Anne Fleur van Luenen
Uppsala University/TNO
annefleur.vanluenen.5513@student.uu.se

Abstract

Studies seeking to recognise moral foundations in written texts have been relatively successful. However, there are two issues with these studies: firstly, it is an extensive process to gather and annotate sufficient material for training. Secondly, models are only trained and tested on different sets of the same data (in-domain). It is yet unexplored how these models perform when tested in other domains. The aim of this study was to test the performance of out-domain moral foundations classification, specifically on extremist data, as it would be useful information when creating a counter narrative. We found that out-domain classification is indeed possible, but with a sharp decline in accuracy. We compared the performance of two models using different types of word embeddings on both in- and out-domain classification. This resulted in a comparable performance for the two models. Finally, we suggest an approach that would be worth exploring further.

1 Introduction

In a series of papers, Haidt and Graham have developed the idea of moral foundations (Haidt and Graham, 2009; Graham et al., 2009; Haidt, 2012; Graham et al., 2013). The idea is that all of our ideas about good and bad are based on five dichotomies. These so-called moral foundations are described in Table 1. Moral foundations are universal, but people differ in how they rank the moral foundations in terms of importance. For example, some people believe an extensive social welfare system is important, because everyone should be able to afford basic necessities. This argument is based on the moral foundation care. Other people believe a social welfare system should be less extensive, because one should be rewarded proportionally to what they contribute to society. This argument is based on fairness. Both people agree that we should take care of people, and that the welfare system should be fair. What differs is which moral foundation they find more important. This gives us insight in people's motivations, and also shows for which moral foundations people are susceptible. This is relevant in the context of extremism, because understanding the moral rhetoric of extremist groups helps us to create a suitable counter narrative.

Moral Foundation	Explanation
Care/harm	Compassion and nursing
Fairness/cheating	Honesty and justice
Loyalty/betrayal	Solidarity and bigotry
Authority/subversion	Obedience and tradition
Sanctity/degradation	Hygiene and purity

Table 1: The five dichotomies of morality according to Graham et al. (2013)

Practical applications require one moral foundations recognition model that is applicable to all types of extremist discourse (i.a. right extremist, separatist groups, etc.). Besides, annotating data is very expensive. Therefore, it would be convenient if a model trained on one corpus could generalise to other types of data. For training, we used the available Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2017), which consists of roughly 35,000 English language tweets on various topics that have been annotated for moral foundations. For testing we collected and annotated our own data (see Section 3)

In our experiment we tested to which extent models trained to recognise moral foundations in non-extremist data generalised to the out-domain (extremist data). Secondly, we compared performance of Word2Vec embeddings, which have been previously used by Lin et al. (2018), to the newer BERT embeddings.

2 Previous Work

Automatic moral foundation recognition started with word count methods. Graham et al. (2009) compiled a dictionary (the Moral Foundation Dictionary, MFD) with words related to each moral foundation. Whenever a word was present in the text, the corresponding moral foundation would be counted.

In the rest of this section we will discuss a number of more advanced models for automatic moral foundation recognition. In all studies, the five dichotomies are split into ten separate moral foundations. Although they are ideologically related, the two parts of the dichtonomy are often discussed in completely different terms, e.g. showing respect and following orders (authority) versus protesting and criticizing (subversion). This results in an eleven-class classification problem: the ten moral foundations, and one 'non-moral' category for samples that do not express moral ideas.

Mooijman et al. (2017) manually labelled 5000 tweets on the Baltimore Protests, which were later included in the MFTC. These tweets were converted into word embeddings, which were passed on to a Long Short Term Memory network (LSTM). The LSTM output vector was concatenated with other dense features, such as the percentage of words that match each category in the MFD. Separate models were trained for all moral foundations and the results were combined. This resulted in an accuracy of 0.890 and an F1-score of 0.880.

Rezapour et al. (2019) used an LSTM containing the word embeddings of the inputdata as a baseline model. They created a variety of other models which used the MFD as secondary input. The best performing model was the one where the words from the MFD were POS-tagged, to ensure homonyms were not counted if their meaning were unrelated to moral foundations. For example, *safe* (adjective) is included, but *safe* (noun) is not. This model obtained an accuracy of 0.866.

Lin et al. (2018) used not only the sample's Word2Vec embeddings, but also a background knowledge vector: a numerical representation of the term frequency of key words in a relevant Wikipedia page abstract. The word embeddings and the background vectors were loaded into separate LSTMs, which were concatenated before they were passed on to the rest of the model. The F1-score fluctuated between the moral foundations, with the lowest score for purity (0.374) and the highest score for care (0.823). Critically, the addition of the background knowledge vector did improve the model.

3 Data

The aim of this study was to train a model on non-extremist data, that recognised moral foundations in extremist data. This was an out-domain task. However, our extremist test data differed from the non-extremist training data in another way than (non-) extremism: the training data consisted of tweets while the test data consisted of forum messages. In order to accurately estimate what part of the performance difference was caused by the test data being extremist, we used three test sets: 1) a subset of the MFTC, consisting of non-extremist tweets, 2) Stormfront data, consisting of extremist forum messages, and 3) Reddit data, consisting of non-extremist forum data. For an overview of the properties of each dataset, see Table 2. We will describe each data set in more detail below.

Data	Extremist or not	Medium
MFTC (test)	non-extremist	Twitter
Reddit	non-extremist	forum
Stormfront	extremist	forum

Table 2: The test sets and their properties

The MFTC was compiled and annotated by Hoover et al. (2017). It consists of 35,000 tweets, of which we were able to scrape 19,022. The tweets concerned six morally relevant topics[1]. They were annotated for the 10 moral foundations and one 1 'non-moral' class. As all data was annotated by three or four annotators, we used the annotation as chosen by the majority. The obtained inter-annotator agreement was a Kappa (Fleiss, 1971) of 0.315 and a Prevalence And Bias Adjusted Kappa (PABAK) (Sim and Wright, 2005) of 0.366. We used 70% of the MFTC for training, 15% for validating the model, and 15% for testing.

Our extremist dataset consisted of a scrape of the Stormfront forum[2], which has been scraped by TNO[3]. Stormfront is described on Wikipedia as "a white nationalist, white supremacist and neo-Nazi,

[1] All Lives Matter; Black Livers Matter; Baltimore protests; 2016 presidential elections; hurricane Sandy; and #MeToo

[2] https://www.stormfront.org/forum/, Note: the content of this forum may be perceived as offensive

[3] See acknowledgements

Internet message forum"[4]. We selected forum posts that contained words related to the six topics of the MFTC, cleaned them, and split them per paragraph. We then selected paragraphs with a length similar to Twitter messages (<300 characters).

Reddit[5] is a forum containing a wide variety of topics, including sports, hobbies, politics and academia. Although extremist opinions can be found on Reddit, the vast majority is non-extremist, making it a useful counterpart for the Stormfront corpus. We gathered the Reddit data by ourselves, by scraping the first 1000 posts that could be found using keywords related to the MFTC topics. Used paragraphs were selected in a similar way compared to the Stormfront paragraphs.

Annotators were recruited among friends and colleagues. All were below 29 years of age and received University-level education in the Netherlands. They had no further background in moral psychology. They received training according to Weber et al. (2018)'s moral foundation annotation training[6], although the training was modified to instruct annotation per tweet or forum message, instead of annotation per word. Each sample was annotated by two annotators, resulting in a Kappa of 0.25 and a PABAK of 0.40 for the Stormfront data (429 samples), and a Kappa of 0.29 and a PABAK of 0.51 for the Reddit data (392 samples). In the roughly one third of the cases where the two annotators did not agree about annotation, we made the final decision between the classifications suggested by the annotators. The low inter-annotator agreement is worrisome, but not uncommon in the field of moral foundations as can be seen by the inter-annotator agreement obtained by Hoover et al. (2017). See Weber et al. (2018) for a discussion.

4 Method

In our experiment we explored how well moral foundation recognition models generalise towards other domains, specifically towards extremist data. Additionally, we tested two models which involve different types of word embeddings. Word embeddings are vectors that numerically represent semantic meaning. The first model uses Word2Vec embeddings (Mikolov et al., 2013), which Lin et al. (2018) also used for automatic moral foundation recognition. We call this the Word2Vec model. For this model, we used a set of 300-dimensional word embeddings pretrained on Google News[7]. The second model uses the newer BERT embeddings (Devlin et al., 2019), we will call it the BERT model.

The newer BERT embeddings solve two issues found with Word2vec. OOV words, which were just ignored by Word2Vec, are processed with word piece embeddings (Wu et al., 2016). This means that a term like *firetruck* will be split in its largest recognisable terms *fire* and *truck* if the term itself is not in the BERT vocabulary. This allowed BERT's vocabulary to be a factor 10 smaller than Word2Vec's vocabulary. Besides, homonyms received separate embeddings for each word sense. Finally, BERT should also capture sentence relations better thanks to its attention mechanism. We used the pretrained 768 dimensional embeddings from the Transformers library[8].

The data was preprocessed according to Rezapour et al. (2019)'s guidelines: URLs, usernames, punctuation and numbers were removed. The # sign was removed from hashtags so that just the word itself remained (e.g. *#BLM* becomes *BLM*). Contractions were expanded using the contractions package[9] (e.g. *I've* becomes *I have*), and all text was lowercased. In the forum data we replaced HTML code *&* with *and*. The samples were then turned into embeddings.

We yielded the best results by double-sampling the classes with less than 1000 occurences in the training data.

Keras[10] was used to build our models. The best performing models were found experimentally. The best performing Word2Vec model had the following architecture: the embeddings were used as input to a bidirectional LSTM network of size 200. This was followed by a dense layer of size 777 with Elu activation, and a softmax output layer of size 11. We used categorical crossentropy loss and Nadam optimization with a learning rate of 0.1.

The best performing BERT model on the development set was the following: BERT embeddings are used as input to a bidirectional LSTM consisting of 250 nodes. This was followed by a dropout

[4] https://en.wikipedia.org/wiki/Stormfront

[5] https://www.reddit.com/

[6] Many thanks to Weber et al. (2018) for allowing us to use their training materials

[7] https://code.google.com/archive/p/word2vec/

[8] https://huggingface.co/transformers/model_doc/bert.html

[9] https://pypi.org/project/contractions/

[10] https://keras.io/

Data	BERT	Word2Vec
MFTC (test)	0.56 (±0.02)	0.59 (±0.01)
Reddit	0.52 (±0.03)	0.51 (±0.03)
Stormfront	0.43 (±0.02)	0.42 (±0.02)

Table 3: Accuracy and standard deviation (between brackets) for each test corpus on the two models. Accuracy is averaged over five runs to compensate for random initialisation.

Data	BERT	Word2Vec
MFTC (test)	0.55	0.60
Reddit	0.47	0.48
Stormfront	0.38	0.38

Table 4: F1-score for each test corpus on the two models. F1-score is averaged over five runs to compensate for random initialisation.

layer of size 0.2; a dense layer of 256 nodes with Elu activation; another dropout layer of size 0.4; and an output layer of size 11 with softmax activation function. We used categorical crossentrophy loss and Nadam optimizer with a 0.1 learning rate.

Both models were trained for 20 epochs. Scores were averaged over five runs to compensate for random initialisation.

5 Results

The results are presented in Tables 3 and 4. The differences in performance between the MFTC test set and the Reddit set is relatively small, whereas the difference in performance between the Reddit set and the Stormfront set is large. If we compare the performance of the different models, the performance of the BERT model is worse than the performance of the Word2Vec model on the test set of the MFTC corpus. The BERT model generalises marginally better to the Reddit and Stormfront data in terms of accuracy. In terms of weighted average F1 score, Word2Vec generalises slightly better to the Reddit data, while the generalisation to the Stormfront data is equal.

6 Discussion

The difference in medium (twitter vs forum) is reflected in the difference in results between the MFTC test data and the Reddit data. This seems to affect the performance of the model less than difference in ideology (non-extremist vs extremist), which is reflected in the difference in results between the Reddit and Stormfront data. If we look into the results qualitatively, it is noticeable that both models struggle specifically with the loyalty class in our extremist test set. In non-extremist data, loyalty takes shape as cheering for a football club or nationalism. On Stormfront, it takes shape as racism. The model has had no opportunity to learn about racism, as it was rare in the training data, and therefore does not recognise it in the text data. The performance of the two models on the other classes is proportional to the occurrence of these classes in the training data.

The performance of our model, even on the MFTC test set, is noticeably lower than the performance of previous studies. This is mainly caused by the previous studies training and testing on MFTC topics separately. This shows that even in in-domain classification, moral foundations classification is complicated by diversity in the data.

As BERT obtained 'state-of-the-art' results on various task sets when it was released (Devlin et al., 2019), it is surprising that the BERT model does not convincingly outperform the Word2Vec model on this task. Multi-word hash tags (e.g. '#blacklivesmatter') are captured by BERT whereas they are regarded as unknown to Word2Vec. Nevertheless, we find comparable numbers of these hash tags among the wrongly classified samples for each model. Any differences in the numbers we find could easily be attributed to other factors as well. We also find no clear differences in the number of homonyms among the wrongly classified samples. Therefore the question why BERT does not outperform Word2Vec remains.

A possible way forward is by extending Lin et al. (2018)'s entity linking approach. This requires Named Entity Recognition of each sample, looking up these samples on Wikipedia and extracting its respective Wikipedia page abstract. The abstracts of all entities in a sample would be turned into a word embedding, which could be added as a secondary input to the model. This background information could be extremely valuable in case of generalising to extremist data, as it would allow to associate less famous extremists to more famous extremists that might occur in the training data.

The key finding of this paper is that moral foundations recognition is complicated by any type of diversity: both in training and test data. We show that out-domain classification is possible, but with the current model it still results in a sharp decline in accuracy.

Acknowledgements

This study describes work done in the master's thesis course at Uppsala University's master in Language Technology. It was carried out as part of the project Opponent Modelling (OM) at TNO, the Dutch Research and Technology institute. Within OM, the data science team aims to extract information from online news sources and fora, in order to gain insight in different opposing groups.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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

Jonathan Haidt and Jesse Graham. 2009. Planet of the durkheimians, where community, authority, and sacredness are foundations of morality. *Social and psychological bases of ideology and system justification*, pages 371–401.

Joseph Hoover, Kate Johnson-Grey, Morteza Dehghani, and Jesse Graham. 2017. Moral values coding guide.

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. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Marlon Mooijman, Joseph Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2017. When protests turn violent: The roles of moralization and moral convergence. *PsyArXiv:4bvyx*.

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.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

René Weber, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2018. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, 12(2-3):119–139.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.