Tomea: an Explainable Method for Comparing Morality Classifiers across Domains

Enrico Liscio[1], Oscar Araque[2], Lorenzo Gatti[3], Ionut Constantinescu[4], Catholijn M. Jonker[1,6], Kyriaki Kalimeri[5], and Pradeep K. Murukannaiah[1]

[1] TU Delft, Delft, the Netherlands
[2] Universidad Politécnica de Madrid, Madrid, Spain
[3] University of Twente, Enschede, the Netherlands
[4] ETH Zürich, Zürich, Switzerland
[5] ISI Foundation, Turin, Italy
[6] Leiden University, Leiden, the Netherlands

1 Introduction

Morality guides us in discerning right from wrong. Language is how we express moral rhetoric, and supervised classification models have shown the ability to recognize moral language in text [1, 2, 18]. However, moral expressions are influenced by *context* [3, 9], which is composed of factors such as actors, actions, and values [22]. For a text classifier, the *domain* in which the data is collected constitutes the context, with recent works [11, 14] analyzing the out-of-domain performance of morality classifiers. Yet, the supervised learning paradigm can lead to black-box models [5], and what causes classifiers to perform differently across domains has not been systematically explored. Such insight is essential to understand whether language models can learn a domain-specific representation of morality, which is especially crucial in delicate applications like healthcare [4].

Contribution In this extended abstract, we summarize our work published at ACL [13]. Our contribution is two-fold. (1) We propose Tomea, an explainable AI (XAI) method to compare a text classifier's representation of morality across domains. Tomea generates domain-specific *moral lexicons* that enable both a quantitative and qualitative comparison of the linguistic cues that a text classifier considers for detecting moral language across domains. (2) We employ Tomea to compare moral rhetoric across seven social domains. We evaluate Tomea through a crowdsourced study involving 159 annotators and by correlating its results to the out-of-domain performance of the employed text classifier.

2 Method

Tomea compares a classifier's representation of morality across domains. Tomea takes as input two ⟨dataset, classifier⟩ pairs, where, in each pair, the classifier is trained on the corresponding dataset (with the datasets assumed to be collected in different domains). Tomea produces a qualitative and quantitative representation of the differences in moral expressions between the two domains.

First, we use an XAI method, SHAP [19], to generate *moral lexicons*, sets of words that describe the classifiers' interpretable representations of the different moral elements in the analyzed domain. For each word in the dataset and for each predicted moral element, SHAP attributes an importance that is proportional to how relevant the model considers the word for predicting that moral element. We compile a moral lexicon for each moral element in a domain, and refer to the union of all the moral lexicons generated in a domain as *domain lexicon*.

Second, we compare the moral and domain lexicons generated for the two domains. We compare the moral lexicons by computing an m-distance for each moral element as the normalized Euclidean distance between the importances determined in the two domains for the words in common between the two moral lexicons. Then, we compare two domain lexicons by computing one d-distance as the Euclidean norm of all the m-distances computed between the two domains.

3 Results and Discussion

We test Tomea on the Moral Foundation Twitter Corpus [10], composed of over 35k tweets collected in seven domains, ranging from MeToo to Hurricane Sandy. Each tweet is annotated with the moral elements of the Moral Foundations Theory [8], which postulates that morality can be deconstructed into 10 irreducible moral elements. We train a multi-label BERT [6] model on each of the seven domains, and employ Tomea to perform pairwise comparisons across the domains.

First, we perform a crowd study where we measure a moderate positive correlation between Tomea's m-distances and human judgment, showing that Tomea can quantify the differences in how the moral elements are represented across domains. Then, we measure a strong negative correlation between d-distances and out-of-domain performance of the model, showing that, the lower the d-distance between two domains, the higher the chance that a model trained on one domain has a good classification performance on the other domain, and vice versa.

In addition to the quantitative analyses, Tomea enables detailed qualitative comparisons of moral expressions across domains. For instance, Tomea deems All Lives Matter (ALM) and Black Lives Matter (BLM) (two of the seven analyzed domains) generally similar. However, the m-distance for the *subversion* moral element is relatively higher than others. Exploring further, we discover that, for *subversion*, words such as 'overthrow' and 'mayhem' have a high impact in ALM, as opposed to words such as 'encourage' and 'defiance' in BLM, in line with the common intuition that *subversion* is viewed differently in ALM and BLM.

We must understand how language models represent human morality across domains before deploying advanced AI systems in our society. Tomea can be a valuable tool for this purpose. It can be used to investigate how a model interprets moral expressions across situational and temporal dimensions, and across different types of moral values [16, 17]. Tomea can support societal applications such as modeling stakeholders' preferences on societal issues [12, 15, 23, 24], analyzing the impact of the use of renewable technologies [25], identifying arguments in personal experiences reports [7, 20], and predicting violent protests [21].

References

1. Alshomary, M., Baff, R.E., Gurcke, T., Wachsmuth, H.: The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. pp. 8782–8797. ACL '22, ACL, Dublin, Ireland (2022)
2. Araque, O., Gatti, L., Kalimeri, K.: MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowledge-Based Systems **191**, 1–11 (2020)
3. Brännmark, J.: Moral disunitarianism. The Philosophical Quarterly **66**(264), 481–499 (2015)
4. Carriere, J., Shafi, H., Brehon, K., Pohar Manhas, K., Churchill, K., Ho, C., Tavakoli, M.: Case Report: Utilizing AI and NLP to Assist with Healthcare and Rehabilitation During the COVID-19 Pandemic. Frontiers in Artificial Intelligence **4**(2), 1–7 (2021)
5. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A Survey of the State of Explainable AI for Natural Language Processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. p. 447–459. AACL '20, Suzhou, China (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. p. 4171–4186. NAACL '19 (2019)
7. Falk, N., Lapesa, G.: Reports of personal experiences and stories in argumentation: datasets and analysis. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5530–5553. ACL '22, ACL, Dublin, Ireland (2022)
8. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In: Advances in Experimental Social Psychology, vol. 47, pp. 55–130. Elsevier, Amsterdam, the Netherlands (2013)
9. Hill, P.L., Lapsley, D.K.: Persons and situations in the moral domain. Journal of Research in Personality **43**(2), 245–246 (2009)
10. Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A.M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T.E., Chin, J., Leong, C., Leung, J.Y., Mirinjian, A., Dehghani, M.: Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. Social Psychological and Personality Science **11**(8), 1057–1071 (2020)
11. Huang, X., Wormley, A., Cohen, A.: Learning to Adapt Domain Shifts of Moral Values via Instance Weighting. In: Proceedings of the 33rd ACM Conference on Hypertext and Social Media. pp. 121–131. HT '22, ACM (2022)
12. Lera-Leri, R., Bistaffa, F., Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.: Towards Pluralistic Value Alignment: Aggregating Value Systems through l-Regression. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. pp. 780–788. AAMAS '22, IFAAMAS, Online (2022)
13. Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C.M., Kalimeri, K., Murukannaiah, P.K.: What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In: Proceed-

ings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers. pp. 14113–14132. ACL '23, ACL, Toronto, Canada (2023)
14. Liscio, E., Dondera, A.E., Geadau, A., Jonker, C.M., Murukannaiah, P.K.: Cross-Domain Classification of Moral Values. In: Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 2727–2745. NAACL '22, ACL, Seattle, USA (2022)
15. Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R.I., Jonker, C.M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Murukannaiah, P.K.: Value inference in sociotechnical systems. In: Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems. pp. 1774–1780. AAMAS '23, IFAAMAS, London, United Kingdom (2023)
16. Liscio, E., van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K.: Axies: Identifying and Evaluating Context-Specific Values. In: Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems. pp. 799–808. AAMAS '21, IFAAMAS, Online (2021)
17. Liscio, E., van der Meer, M., Siebert, L.C., Jonker, C.M., Murukannaiah, P.K.: What Values Should an Agent Align With? Autonomous Agents and Multi-Agent Systems **36**(23), 32 (2022)
18. Lourie, N., Le Bras, R., Bhagavatula, C., Choi, Y.: UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. pp. 13480–13488. AAAI '21 (2021)
19. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. pp. 1208–1217. NeurIPS '17, Long Beach, CA, USA (2017)
20. van der Meer, M., Liscio, E., Jonker, C.M., Plaat, A., Vossen, P., Murukannaiah, P.K.: HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In: HHAI2022: Augmenting Human Intellect. pp. 17–31. HHAI '22, IOS Press, Amsterdam, the Netherlands (2022)
21. Mooijman, M., Hoover, J., Lin, Y., Ji, H., Dehghani, M.: Moralization in social networks and the emergence of violence during protests. Nature Human Behaviour **2**(6), 389–396 (2018)
22. Schein, C.: The Importance of Context in Moral Judgments. Perspectives on Psychological Science **15**(2), 207–215 (2020)
23. Shortall, R., Itten, A., van der Meer, M., Murukannaiah, P.K., Jonker, C.M.: Reason against the machine? future directions for mass online deliberation. Frontiers in Political Science **4**, 1–17 (10 2022)
24. Siebert, L.C., Liscio, E., Murukannaiah, P.K., Kaptein, L., Spruit, S.L., van den Hoven, J., Jonker, C.M.: Estimating Value Preferences in a Hybrid Participatory System. In: HHAI2022: Augmenting Human Intellect. pp. 114–127. IOS Press, Amsterdam, the Netherlands (2022)
25. de Wildt, T.E., van de Poel, I.R., Chappin, E.J.L.: Tracing long-term value change in (energy) technologies: Opportunities of probabilistic topic models using large data sets. Science, Technology, & Human Values **47**(3), 429–458 (2022)