26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Quantifying Annotator Reliability in Moral Classification of Texts

Yi Ren [a]

[a]*School of Computer and Mathematical Sciences, University of Adelaide*
Email: yi.ren@adelaide.edu.au

Abstract: Moral values and stances play a vital role in decision making, social interaction, opinion sharing and formation of judgements around social issues. The rapid growth of social media has expanded platforms for people to engage and interact with each other in ways that shape and change their moral perspectives. This provides opportunities for social scientists to observe and analyse moral expressions in large-scale online discourse. Moral Foundation Theory (MFT) is a well-known psychological theory that decomposes human moral reasoning into several universal foundations, and can be used to gain insight into how moral values shape decision-making and judgements. Analysing social media data through the lens of MFT offers valuable opportunities to gain a deeper understanding of the moral reasoning that underlies people's behaviour,

Due to the large scale of the data from social media, it is impractical to rely solely on humans to manually extract moral values. Instead, researchers have turned to natural language processing techniques to automatically detect moral values. Supervised classification models trained on human-annotated datasets to extract moral values from texts have been widely adopted in previous work. Specifically, by fine-tuning pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT), one can achieve state-of-the-art classification performance, often outperforming human annotators.

Nonetheless, critical problems in the datasets are often overlooked. Substantial disagreements between annotators in training datasets highlight both the subjective nature of the task and the influence of human biases. Furthermore, training datasets are often collected from a variety of social or political topics, leading to significant differences in underlying moral value distributions across all texts. Existing work typically uses simple techniques such as majority voting to aggregate "true" labels from annotations. However, such heuristics presume annotators are equally reliable and independent; in reality, reliability and bias vary by annotators and topics. These limitations undermine the reliability of training datasets and complicate model evaluation.

Our proposed framework extends the BERT model by including data-aware neural network structures for annotators and topics. We implement two distinct designs of annotator layers, providing different measurements of annotator reliability. The first design estimates the annotators' biases towards moral labels, whereas the second estimates the full confusion matrix of each annotator's labelling accuracy with respect to the hidden ground truth. The topic layer summarises the underlying distributions of moral values in the various topics. While accounting for individual influence of annotators and topics, the model also learns the interacting effect of the two through backpropagation. This talk will explore the performance of our method relative to existing approaches, and highlight the advantages and weaknesses.

Keywords: *Natural Language Processing, Interpretable Neural Network, Computational Social Science*