Explaining NLP Classification of Human Values

Values are abstract motivations that justify opinions and actions (Schwartz 2012). To be societally beneficial, AI should not be morally neutral, but actively strive to align with humans' social goals and interests (Russell et al. 2015). Understanding values is an essential milestone in achieving beneficial AI with applications in fields such as autonomous driving, healthcare, and AI-assisted policy making.

Existing methods for value elicitation typically rely on user surveys (Schwartz, 2012). However, in practical applications (e.g., to conduct meaningful conversations or to identify online trends), artificial agents should be able to identify values on the fly. The growing capabilities of **natural language processing** (NLP) enable the estimation of values from discourse (Mooijman et al. 2018; Hoover et al. 2020). **Value classifiers** can be used to identify the values underlying a piece of text.

NLP models have been proven to be effective in classifying values in text. However, the most advanced deep learning models typically suffer from the ***black-box*** problem: it is often hard to understand the reasoning behind the models' decisions. Understanding the reasoning of ML models is a problem typically referred to as **explainability** (Danilevsky, 2020). Due to the subjective and abstract nature of values, value classifiers explainability is essential for two reasons: (1) to understand whether the reasoning of the model is in line with our intuition and expectations, and (2) to explain the decisions to end users so as to build trust in the system.

The goal of this project is to provide tools to inspect and explain the decisions of a value classifier. We have an array of different NLP models, ranging from LSTM to BERT, and a dataset composed of 35k tweets annotated with values (Hoover et al. 2020). The aim is to elaborate a methodology to inspect the *global* explainability (i.e., the model's prediction process as a whole, as opposed to the explanation for an individual prediction) of a value classifier, independently of the selected model or dataset.

Desired Skills:

- Basic Python experience
- Basic NLP/ML knowledge

For more information, please send an email to Enrico Liscio (e.liscio@tudelft.nl) and Pradeep Murukannaiah (p.k.murukannaiah@tudelft.nl).

References:

Schwartz, S. H. 2012. “An Overview of the Schwartz Theory of Basic Values.” *Online readings in Psychology and Culture* 2(1):1–20.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. “Research Priorities for robust and beneficial artificial intelligence.” *AI Magazine* 36(4):105–114.

Mooijman, Marlon, et al. "Moralization in social networks and the emergence of violence during protests." *Nature human behaviour* 2.6 (2018): 389-396.

Hoover, Joe, et al. "Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment." *Social Psychological and Personality Science* 11.8 (2020): 1057-1071.

Danilevsky, Marina, et al. "A survey of the state of explainable AI for natural language processing." *arXiv preprint arXiv:2010.00711* (2020).