Human Value Detection by Fine Tuning Embedding Models using Value Definitions

Pushkar Bhardwaj
Department of Computer Science
Indian Institute of Technology Kanpur
pushkarb@cse.iitk.ac.in

***Abstract*—Human values form the basis of argumentation and are inherently reflected in language. An understanding of implied human values in language is essential for chat agents to generate relevant responses as well as ethical and cultural compliance. Existing LLMs are shown to lack this understanding. The objective of the present work is detection of one or more implied human values for long text sequences and their attainment. We propose utilizing descriptions of value labels for definition-assisted labelling and compare it to the existing LLM prompting and fine-tuning approaches. Preliminary results show that this approach appears to offer an advantage compared to the earlier two approaches. Also, the performance is shown to be better than out-of-the-box large-language models, implying the need for more focus on value and sentiment detection for LLMs. Presently, we are working on performing a detailed analysis and further hyperparameter tuning.**

***Index Terms*—Human value detection, llm prompting, fine-tuning, definition-assisted labeling.**

Manuscript received Nov 19, 2024