Identifying the Human Values behind Arguments

Johannes Kiesel
Bauhaus-Universität Weimar
johannes.kiesel@uni-weimar.de

Milad Alshomary
Paderborn University
milad.alshomary@upb.de

Nicolas Handke
Universität Leipzig
nicolas.handke@gmx.de

Xiaoni Cai
Technische Universität München
caix@in.tum.de

Henning Wachsmuth
Paderborn University
henningw@upb.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Abstract

This paper studies the (often implicit) human values behind natural language arguments, such as to *have freedom of thought* or to *be broad-minded*. Values are commonly accepted answers to why some option is desirable in the ethical sense and are thus essential both in real-world argumentation and theoretical argumentation frameworks. However, their large variety has been a major obstacle to modeling them in argument mining. To overcome this obstacle, we contribute an operationalization of human values, namely a multi-level taxonomy with 54 values that is in line with psychological research. Moreover, we provide a dataset of 5270 arguments from four geographical cultures, manually annotated for human values. First experiments with the automatic classification of human values are promising, with F_1-scores up to 0.81 and 0.25 on average.

1 Introduction

How come people disagree on the best course forward in controversial issues, even if they use the same information to form their opinion? A way to get to the bottom of such disagreement is to repeatedly ask them why they see something as desirable. We observe that people have different beliefs and priorities of what is generally worth striving for (e.g., personal achievements vs. humility) and how to do so (e.g., being self-directed vs. respecting traditions), often referred to as *(human) values* (Searle, 2003). Some values tend to conflict and others to align (see Figure 1), which can cause disagreement on the best course forward, but also the support, if not formation, of political parties that promote the respective highly revered values. Moreover, one can observe different value priorities between cultures and disagreement thereon.

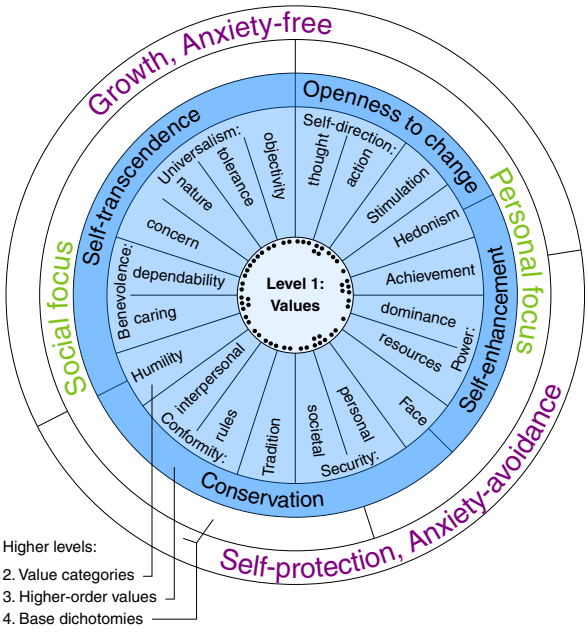

Figure 1: The levels of this paper's consolidated taxonomy of 54 values (shown as black dots) that are categorized on the more abstract levels 2–4 (cf. Section 3). Categories that tend to conflict are placed on opposite sites. Illustration adapted from (Schwartz et al., 2012).

Due to their outlined importance, human values are studied both in the social sciences (Schwartz, 1994) and in formal argumentation (Bench-Capon, 2003) for decades. According to the social sciences, a "value is a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities." As Schwartz continues, these features "make it possible to conclude that security and independence are values,

whereas thirst and a preference for blue ties are not." Consider the following example:

"*Social media is good for us. Though it might make people less polite, it makes our lives much easier.*"

To understand the pragmatics of this argument, a reader has to acknowledge the belief (Point 1 in the definition above) that the "end state" (2) of *having a comfortable life* is desirable in general (3). To concur with the statement (4), the reader further has to prefer *having a comfortable life* over *being polite* (5)—ignoring other arguments on the topic for the sake of the example. Within computational linguistics, human values thus provide the context to categorize, compare, and evaluate argumentative statements, creating several possibilities: to inform social science research on values through large-scale datasets; to assess argumentation with respect to scope and strength; to generate or select arguments based on the value system of a target audience; and to identify opposing and shared values on both sides of a controversial topic.

However, the task to identify values in arguments seems daunting due to their large number, often implicit use in arguments, and vague definitions. On the other hand, the creation of larger argumentation datasets, advancements in natural language understanding, and the decade-long rigorous taxonomization of values by social scientists has put such an automatic identification within reach.

As a first endeavor on the automatic identification of values in written arguments, this paper makes three contributions: (1) a consolidated multi-level taxonomy of 54 human values taken from four authoritative cross-cultural social science studies (Section 3); (2) a dataset of 5270 arguments from the US (most arguments), Africa, China, and India, each of which manually annotated for all values by three annotators, corresponding to about 850k human judgments (Section 4); and (3) first classification results per taxonomy level, establishing a baseline and revealing promising results both within and across cultures (Section 5).

2 Background

Human values are of concern to most if not to all social sciences (Rokeach, 1973) and have also been integrated into computational frameworks of argumentation (Bench-Capon, 2003). In NLP, values have been analyzed for personality profiling (Maheshwari et al., 2017), but not yet for argument mining, as considered here.

2.1 Values in Social Science

Rokeach (1973) already described the two concepts of (1) a value as a belief pertaining to desirable end states or modes of conduct and (2) a value system as prioritization of values based on cultural, social, and personal factors. These definitions attribute values to persons rather than to objects, facilitating a systematic analysis (Rokeach, 1973). The paper at hand follows these definitions and targets the personal values behind arguments, that is, the values that the arguments, mostly implicitly, resort to.

Several proposed value schemes are domain-independent and hence suited to analyze generic argumentation. Our consolidated value taxonomy (Section 3) is thus based on these schemes. Combining research from anthropology, sociology, philosophy, and psychology, Rokeach (1973) estimates the total number of human values to be fewer than hundreds, and develops a practical survey of 36 values that distinguishes between values pertaining to desirable end states and desirable behavior.

Specifically for cross-cultural analysis, Schwartz et al. (2012) derived 48 value questions from the universal needs of individuals and societies, including *obeying all the laws* and *to be humble*. Moreover, Schwartz (1994) proposes a relatedness of values by their tendency to be compatible in their pursuit (see Figure 1). This relatedness reflects two "higher order" conflicts: (1) openness to change/own thoughts vs. conservation/submission, and (2) self-transcension (directed towards others/the environment) vs. self-enhancing (directed towards one's self), allowing to analyse values at several levels. Cheng and Fleischmann (2010) consolidates 12 schemes into a "meta-inventory" with 16 values, such as *honesty* and *justice*, revealing a large overlap in schemes across fields of research. However, as the meta-inventory is strictly more coarse-grained than Schwartz et al.'s theory we do not investigate it further in this paper.

Other schemes, however, pertain to specific purposes, making them less suited for our study. We give an overview for completeness. England (1967) suggested 66 values related to management decisions, such as *high productivity* and *prestige*, and categorized them by relevant entity, for example *business organizations* and *individuals*. Brown and Crace (2002) looked at 14 values for counseling and therapy, such as *responsibility* and *spirituality*, and Kahle et al. (1988) at nine for consumer research, such as *warm relationships* and *excitement*.

2.2 Values in Argumentation Research

Formal argumentation employs value systems to model audience-specific preferences, that is, an argument's strength depends on the degree to which the audience reveres the values the argument resorts to. Examples include value-based argumentation schemes (van der Weide et al., 2009), defeasible logic programming (Teze et al., 2019), and the value-based argumentation framework of Bench-Capon (2003). The latter is an extension of the abstract argumentation framework of Dung (1995) that has already been applied manually to analyze interactions with reasoning and persuasion subject to a specific value system (Atkinson and Bench-Capon, 2021). This paper presents a first step towards the large-scale automatic application of these works as it takes values to argument mining.

Feldman (2021) recently showed the strong connection between values and the moral foundation theory (Haidt, 2012). Like personal values, this theory analyzes ethical reasoning behind human choices, but considers five rather abstract "foundations:" care, fairness, loyalty, authority, and purity. Alshomary and Wachsmuth (2021) hypothesized that the foundations could be used for audience-specific argument generation. Kobbe et al. (2020) tried to classify arguments by foundations, but noted a low human agreement due to the vagueness of the foundations. We assume values can here contribute to the classification by foundations.

Values overlap with idea of *framing* in communication, that is, the selection and emphasis of specific aspects of (perceived) reality to promote a particular problem, causal interpretation, ethical evaluation, and/or recommendation (Entman, 1993). In frames, values can define the costs and benefits of options (Entman, 1993), while common value systems are used for evaluation. Framing has often been studied computationally for news (Naderi and Hirst, 2015; Chen et al., 2021), but also for political speech (De Vreese, 2005), and argumentation (Ajjour et al., 2019). In the latter, some values are so prevalent that they constitute frames of their own, indicating a potential use of values in frame identification. For example, 14 out of 54 values we use are also frames in the dataset of Ajjour et al.[1]

Values may be considered as aspects under which to group arguments. Some researchers have mined aspects from text (Trautmann, 2020) or used them to control argument generation (Schiller et al., 2021). Others have studied the task of opinion summarization in arguments (Egan et al., 2016; Misra et al., 2016; Chen et al., 2019), aiming at the most important aspects discussed in a debate. Related, the task of key point analysis (Bar-Haim et al., 2020; Friedman et al., 2021) is to generate a small set of concise statements that each represent a different aspect. We argue that analyzing the values found in a collection of arguments provides a new perspective to aspects in argumentation, focusing on the "why" behind an argument's reasoning.

3 Taking Values to Argument Mining

Human values have been considered in formal argumentation since about 20 years (Bench-Capon, 2003). However, to the best of our knowledge, our paper is the first that aims at identifying the values behind arguments computationally. The term "behind" reflects the fact that many arguments do not explicate values; for example, in the argument "no matter they felt forced to commit it: anyone who commits a crime should be prosecuted" no value is mentioned literally. The argument gains its persuasive strength when being connected to values, which can be both desirable behavior (*behaving properly*) or end states (*a safe country*). By putting forward an argument, its proponent wants the audience to connect the argument with its values. Formally, values are connected specifically with the argument's premise. However, automatic models might still improve when incorporating the textual conclusion as context for the textual premise. The task studied in this paper is to draw this connection between arguments and values automatically.

The heart of a value-based argumentation framework is a value taxonomy (or a set of values) that is both accepted and relevant. The research presented in this paper is largely based on the refined theory of Schwartz et al. (2012),[2] which, however, has been extended by us: Comparing Schwartz et al.'s refined theory with three other widespread value lists against a sample of our dataset, we decided to add and integrate nine values (see Table 1). We also asked the annotators to comment on supposedly missing values (see Section 4). For most of the additional 48 value descriptions that we received (*be humane*, *be fair*, *be modern*, etc.), we identified existing values or value combinations in the taxonomy that subsume them, suggesting to extend the value descriptions rather than adding new values.

[1] Per Jaccard similarity of value and frame names ≥ 0.5.

[2] Using the noun-phrase value names of Schwartz (1994).

Level				Source				Dataset frequency (size; cf. Section 4)			
4a/4b	3	2) Value category	1) Value	SVS	RVS	LVI	WVS	Africa (50)	China (100)	India (100)	USA (5020)
Growth, Anxiety-free	Openness to change	Self-direction: thought	Be creative	●		○	○	0.000	0.040	0.020	0.028
			Be curious	●				0.000	0.030	0.020	0.049
			Have freedom of thought	●	○		○	0.080	0.000	0.040	0.124
		Self-direction: action	Be choosing own goals	●			○	0.000	0.030	0.040	0.135
			Be independent	●	○		○	0.080	0.030	0.000	0.100
			Have freedom of action	●	○		○	0.080	0.030	0.030	0.171
			Have privacy			○	○	0.000	0.040	0.070	0.019
		Stimulation	Have an exciting life	●	○		○	0.000	0.000	0.010	0.020
			Have a varied life	●				0.000	0.000	0.000	0.041
			Be daring	●				0.000	0.000	0.000	0.010
	Self-enhancement	Hedonism	Have pleasure	●	○		○	0.000	0.020	0.010	0.039
		Achievement	Be ambitious	●	○		○	0.020	0.050	0.050	0.048
			Have success	●			○	0.100	0.160	0.120	0.127
			Be capable	●	○		○	0.040	0.200	0.150	0.146
			Be intellectual		○		○	0.040	0.130	0.020	0.065
			Be courageous		○			0.020	0.000	0.000	0.009
		Power: dominance	Have influence	●			○	0.040	0.010	0.000	0.057
			Have the right to command	●			○	0.000	0.000	0.010	0.042
		Power: resources	Have wealth	●			○	0.060	0.190	0.030	0.108
Self-protection, Anxiety-avoidance	Conservation	Face	Have social recognition	●	○			0.040	0.000	0.020	0.050
			Have a good reputation	●				0.020	0.010	0.030	0.026
		Security: personal	Have a sense of belonging	●		○		0.100	0.010	0.020	0.081
			Have good health	●			○	0.080	0.030	0.120	0.123
			Have no debts	●				0.000	0.020	0.020	0.051
			Be neat and tidy	●	○			0.000	0.000	0.000	0.002
			Have a comfortable life		○		○	0.080	0.260	0.190	0.199
		Security: societal	Have a safe country	●	○		○	0.160	0.030	0.180	0.183
			Have a stable society	●			○	0.420	0.300	0.170	0.228
		Tradition	Be respecting traditions	●			○	0.020	0.000	0.020	0.089
			Be holding religious faith	●			○	0.000	0.000	0.050	0.052
		Conformity: rules	Be compliant	●	○		○	0.040	0.070	0.100	0.136
			Be self-disciplined	●	○			0.000	0.030	0.010	0.029
			Be behaving properly			○	○	0.160	0.070	0.180	0.147
		Conformity: interpersonal	Be polite	●	○		○	0.000	0.010	0.030	0.031
			Be honoring elders	●			○	0.000	0.000	0.000	0.012
		Humility	Be humble	●			○	0.080	0.020	0.010	0.014
			Have life accepted as is	●				0.040	0.040	0.040	0.074
Personal focus	Self-transcendence	Benevolence: caring	Be helpful	●	○		○	0.060	0.030	0.040	0.155
			Be honest	●	○		○	0.060	0.010	0.020	0.045
			Be forgiving	●	○			0.000	0.000	0.010	0.019
			Have the own family secured		○		○	0.000	0.090	0.030	0.083
			Be loving		○		○	0.020	0.020	0.040	0.054
		Benevolence: dependability	Be responsible	●	○		○	0.060	0.030	0.110	0.146
			Have loyalty towards friends	●			○	0.000	0.000	0.000	0.003
		Universalism: concern	Have equality	●	○	○	○	0.240	0.090	0.200	0.165
			Be just	●			○	0.060	0.180	0.160	0.251
			Have a world at peace	●	○		○	0.260	0.000	0.040	0.091
		Universalism: nature	Be protecting the environment	●			○	0.000	0.080	0.010	0.036
			Have harmony with nature	●				0.000	0.050	0.050	0.055
			Have a world of beauty	●	○		○	0.000	0.000	0.000	0.012
		Universalism: tolerance	Be broadminded	●	○		○	0.100	0.010	0.090	0.102
			Have the wisdom to accept others	●	○		○	0.020	0.010	0.000	0.059
		Universalism: objectivity	Be logical		○		○	0.020	0.120	0.090	0.082
			Have an objective view			○	○	0.100	0.160	0.100	0.126

Table 1: The 54 values of the taxonomy with sources and dataset frequency. Level 4a contains two labels, *personal focus* and *social focus* while 4b refers to motivation regarding anxiety. Following Schwartz et al. (2012), each value has one label per level, except *have pleasure* (both *self-enhancement* and *openness to change* for Level 3) and the achievement values (both Level 4b labels). The main source taxonomy (●) is the Schwartz Value Survey (SVS, Schwartz et al., 2012). Additional values are taken from (○) the Rokeach Value Survey (RVS, Rokeach, 1973), the Life Values Inventory (LVI, Brown and Crace, 2002), and the World Values Survey (WVS, Haerpfer et al., 2020).

Only two of the added values are not directly related to the universal needs that Schwartz (1994) based the value categories on. The proposed category *universalism: objectivity* integrates well between the outward thinking of *universalism: tolerance* and the free thinking of *self-direction: thought* (see Figure 1). We adopt a uniform naming scheme where the value names reflect the distinction of Rokeach (1973) into instrumental (*be* ...) and terminal (*have* ...) values, and are easy to embed in sentences, for example, "it is good to *be creative*."

The taxonomy levels are chosen based on usefulness in social science research. The values at Level 1 are intended to be the items in surveys (Schwartz, 1994), which is why we also suggest to use them for dataset annotation. Moreover, Level 1 values can still be classified into being either instrumental or terminal. One could, however, create arbitrarily coarse- and fine-grained levels.[3]

The close connection of our taxonomy to social science research enables studies of value systems across disciplines that are beyond the scope of this paper. The grouping of values at higher levels allows for classifications at coarser levels of granularity, enabling investigations such as, whether a specific set of arguments focus on persons or society mainly, or whether they imply a rather anxiety-free or a rather anxiety-avoiding background (cf. Figure 1). Also, the circular organization of the taxonomy enables the analysis of major "directions" in a collection of arguments, which can, for example, be used to study value differences in argumentation datasets of different cultures. In addition, for the 41 values with a link to the World Values Survey (the WVS column in Table 1, Haerpfer et al., 2020), the corresponding dataset contains information on people's value priorities (i.e., value systems) collected rigorously for 51 territories, with the earliest survey from 1981 and the latest from 2020. These links allow comparing value distributions identified in regional datasets with survey data.

4 A Dataset of Values behind Arguments

This section presents the first dataset for studying human values behind arguments. Each of the 5270 arguments included was annotated by three crowdworkers for all 54 values from Section 3. The dataset, taxonomy description, and annotation interface are available online as Webis-ArgValues-22.[4]

[3] For example, with values such as "have no broken legs".

[4] https://github.com/webis-de/ACL-22

4.1 Argument Sources of Different Cultures

Following the aspiration of a cross-cultural value taxonomy and using territories as a proxy for cultures, the dataset is composed of four parts: *Africa*, *China*, *India*, and *USA*. Each argument consists of one premise, one conclusion, and a stance attribute indicating whether the premise is in favor of (pro) or against (con) the conclusion. As existing argument datasets are almost exclusively from a Western background, we had to collect new suitable arguments for the non-US parts, drastically limiting their size. The respective non-US sources were recommended to us for their authenticity by students from the respective territory that work with our groups. Note that this data is not intended to represent the respective culture, but to train and benchmark classifiers across sources.

Africa We manually extracted 50 arguments from recent editorials of the debating ideas section of a pan-African news platform, *African Arguments*.[5] Premises could often be extracted literally, but conclusions were mostly implicit and had to be compiled from several source sentences.

China We extracted 100 arguments from the recommendation and hotlist section of a Chinese question-answering website, *Zhihu*.[6] We manually identified key points (premises and conclusions) in the answers and manually translated them to English using automated translation for a first draft.

India We extracted 100 arguments from the controversial debate topics 2021 section of *Group Discussion Ideas*.[7] This blog collects pros and cons on various topics from Indian news to support discussions. Premises and conclusions were used as-is.

USA We took 5020 arguments with a manual argument quality rating of at least 0.5 from the 30,497 arguments of the IBM-ArgQ-Rank-30kArgs dataset (Gretz et al., 2020). For the dataset, crowdworkers wrote one pro and one con argument for one of 71 common controversial topics. We rephrased the topics to represent conclusions.

Due to the difficulty of collecting datasets from various cultures, the number of respective arguments (250) is small compared to the US part. However, we will mainly use them for testing the robustness of identifying values in arguments.

[5] https://africanarguments.org

[6] https://www.zhihu.com

[7] https://www.groupdiscussionideas.com

Argument	Values	Dataset part
○ Pro "South Africa's COVID-19 lockdown was too strict": The economic ramifications of the lockdown have been huge, and have been felt hardest by those who were already most vulnerable.	Have a comfortable life, Have a stable society, Have equality	Africa
○ Pro "We should protect our privacy in the Internet age.": The leaked personal information will be defrauded by fraud gangs to gain trust and carry out fraudulent activities.	Have privacy, Have a stable society, Be compliant	China
○ Con "Rapists should be tortured": Throughout India, many false rape cases are being registered these days. Torturing all of the accused persons causes torture to innocent persons too.	Have a safe country, Have a stable society, Be just	India
○ Pro "We should adopt an austerity regime": An austerity regime will help to reduce the deficit of the country.	Have no debts, Have a stable society, Be responsible	USA

Table 2: Four example arguments (stance, conclusion, and premise) and their annotated values. We selected these to showcase different ways for resorting to *have a stable society*, the most frequent value, from each dataset part.

Part	Conclusions		Premises		Stances	
	#	Tokens	#	Tokens	# Pros	# Cons
Africa	23	10.6	50	28.1	37	13
China	12	7.3	100	24.5	59	41
India	40	6.6	100	30.3	60	40
USA	71	5.6	5020	18.5	2619	2401
Total	146	5.6	5270	18.9	2775	2495

Table 3: Numbers of unique conclusions and premises for each part of the contributed dataset, their mean number of space-separated tokens, and stance distribution.

Table 2 shows one example from each part. Note that we do not see any part as representative for the respective culture, but rather as a necessary approximation (see Section 7 for a discussion). Table 3 provides an overview of the dataset. Premises are longer than conclusions, with USA having the lowest average for both. The Africa part has the fewest premises per conclusion (2.2) and the US part the most (70.7). The skew between pros and cons is highest for Africa with a ratio of about 3:1. All these observations are results of the collection process and are natural variations for arguments.

4.2 Crowdsourcing of Value Annotations

We employed a custom three-part annotation interface, optimized for speed and task expertise acquisition through keyboard shortcuts and a clear template-like structure (see Appendix A for screenshots). Besides instructions and example arguments, a brief explanation of specific terms was given if needed (e.g., for the "996 overtime system" mentioned in several arguments from China). Below this introductory material, the main part of the interface consists of three panels. The first panel places the argument to be annotated in a scenario:

> *Imagine someone is arguing [in favor of/against] "[conclusion]" by saying: "[premise]."*

The second panel formulated the annotation task for a value as a yes/no question.[8] The question follows the operationalization of Section 3:

> *If asked "Why is that good?", might this be their justification? "Because it is good to [value]".*

For illustration, example implications of matching arguments were provided. Instructions stated that one to five values are typical for an argument, and more than 10 should be avoided. A third panel shows the annotation progress. Annotators could write feedback on both arguments and values.

The crowdsourcing ran on the MTurk platform, with annotators taking 2:40 minutes per argument on average, and totaling 90 days of 8-hour work. We required them to have an approval rate of at least 98%, at least 100 approved work tasks, and—for language proficiency—being located in the US. No further personal information was gathered. The annotators were first restricted to three annotation tasks. Manual quality checks at this stage resulted in 154 work rejections (5% rejection rate) due to ignored instructions. We then selected 27 annotators for annotating the bulk of arguments, ensuring at least 3 annotations per argument. As mandatory for MTurk, annotators were paid on a task basis, which led to an average hourly wage of $8.12 (current US federal minimum wage: $7.25). Additionally, we paid bonuses of total $65.65, especially to annotators who wrote extensive comments.

[8] To prevent order effects, the value order was randomized for each annotator, but then fixed to allow for learning.

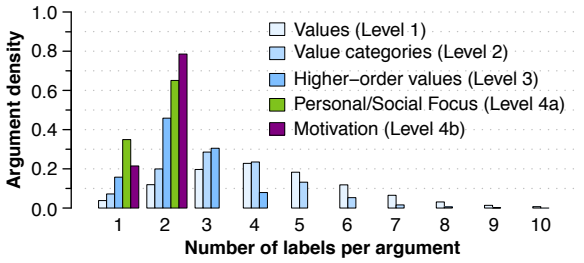

Figure 2: Fraction of arguments having a specific number of assigned labels for each level. The total number of labels for levels 1–4b are 54, 20, 4, 2, and 2.

We employed MACE (Hovy et al., 2013) to fuse the annotations into a single ground truth, applying it value-wise as suggested by the author for multi-label annotations. Despite the difficulty of the annotation task, the crowdworker annotators reached an average value-wise agreement α of 0.49 (Krippendorff, 2004). We found most disagreement arose from the complexity of annotating 54 values at once, with annotators sometimes confusing values despite the descriptions. For follow-up datasets, one could likely reduce such problems by training annotators on the arguments of our dataset with highest disagreement. One step we implemented for quality assurance is that we manually checked the 48 arguments (<1%) to which MACE assigned more than 10 values, reducing their values to the most prevalent 5–7 ones. The right side of Table 1 shows the frequency of each value in each dataset part, revealing that each value occurs at least once.

A value in the ground truth also automatically led to an assignment of all parent labels in the taxonomy (see Figure 1). Figure 2 shows the resulting level-wise distribution of labels per argument. As the majority of arguments are assigned both labels for Levels 4a and b, these base dichotomies for values are hence mostly not dichotomous for arguments. So, like the value systems of people, many arguments seem to resort to a broad spectrum of values from the value continuum at once. For example, the first argument in Table 2 resorts to both *having a comfortable life* (personal focus, self-protection) and *having equality* (social focus, growth). Similar to observations of Rokeach (1973, p. 50f) on value systems, this example showcases an interaction between values that change their psychological significance, where *having equality* gives *having a comfortable life* a social focus. We believe that our dataset enables scholars to study such interactions for arguments in the future.

5 Identifying Values behind Arguments

This section presents a first attempt at automatically identifying human values using standard approaches. The first experiment focuses on the USA dataset part alone, the second on a cross-cultural setting. We compare three approaches, for which we provide our implementation online:[9]

BERT. Fine-tuned multi-label bert-base-uncased with batch size 8 and learning rate 2^{-5} (20 epochs).

SVM. A linear kernel scikit-learn support vector machine trained label-wise with $C = 18$.

1-Baseline. Classifies each argument as resorting to all values. Thus always achieves a recall of 1.

Our evaluation focuses on the label-wise F_1-score and its mean over all labels (macro-average), as well as its constituents precision and recall. We report accuracy for completeness, though the heavily skewed label distribution makes it less suited. The evaluation employs macro-averages for all metrics to give the same weight to all values. Note that the 1-Baseline is especially strong for the F_1-score since it always achieves a recall of 1. By definition this baseline achieves at least as high—and in most cases higher—F_1-scores than label-wise random guessing according to the label frequency. For calculating the p-values when comparing approaches we employ the Wilcoxon signed rank significance test (Wilcox, 1996). As detailed in Section 4, most arguments actually have both labels of the base dichotomies (Levels 4a and b) assigned to them, so we do not discuss these levels deeper here.

5.1 Results on the USA Part

We first report results on the main part of our dataset (USA) as an experiment with matching training and test set. The approaches are trained on the arguments from 60 unique conclusions (4240 arguments, ~85%), validated on 4 (277, ~5%), and tested on 7 (503, ~10%). The conclusions were selected so that the different sets contain roughly the specified percentage of arguments. Unfortunately, this process led to different value distributions in the different sets. However, we deemed the conclusion-wise split more important for our experiments, as we want to test whether classifiers generalize to unseen conclusions. Only one very rare value, *be neat and tidy* (0.2% of arguments in USA part), does not occur in the test set. We thus exclude this value from evaluation.

[9] https://github.com/webis-de/ACL-22

Model	Level 1				Level 2				Level 3				Level 4a				Level 4b			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
BERT	**0.40**	0.19	**0.25**	**0.92**	**0.39**	0.30	**0.34**	**0.84**	0.65	0.78	0.71	**0.67**	**0.89**	0.96	**0.92**	**0.86**	0.92	**1.00**	**0.96**	**0.92**
SVM	0.21	0.19	0.20	0.88	0.30	0.30	0.30	0.77	**0.66**	0.68	0.67	0.65	0.88	0.89	0.88	0.80	**0.93**	0.90	0.92	0.85
1-Baseline	0.08	**1.00**	0.16	0.08	0.18	**1.00**	0.28	0.18	0.60	**1.00**	**0.75**	0.60	0.85	**1.00**	**0.92**	0.85	0.92	**1.00**	**0.96**	**0.92**

Table 4: Macro precision (P), recall (R), F_1-score (F_1), and accuracy (Acc) on the USA test set over all labels by level. Best scores per metric and level marked bold.

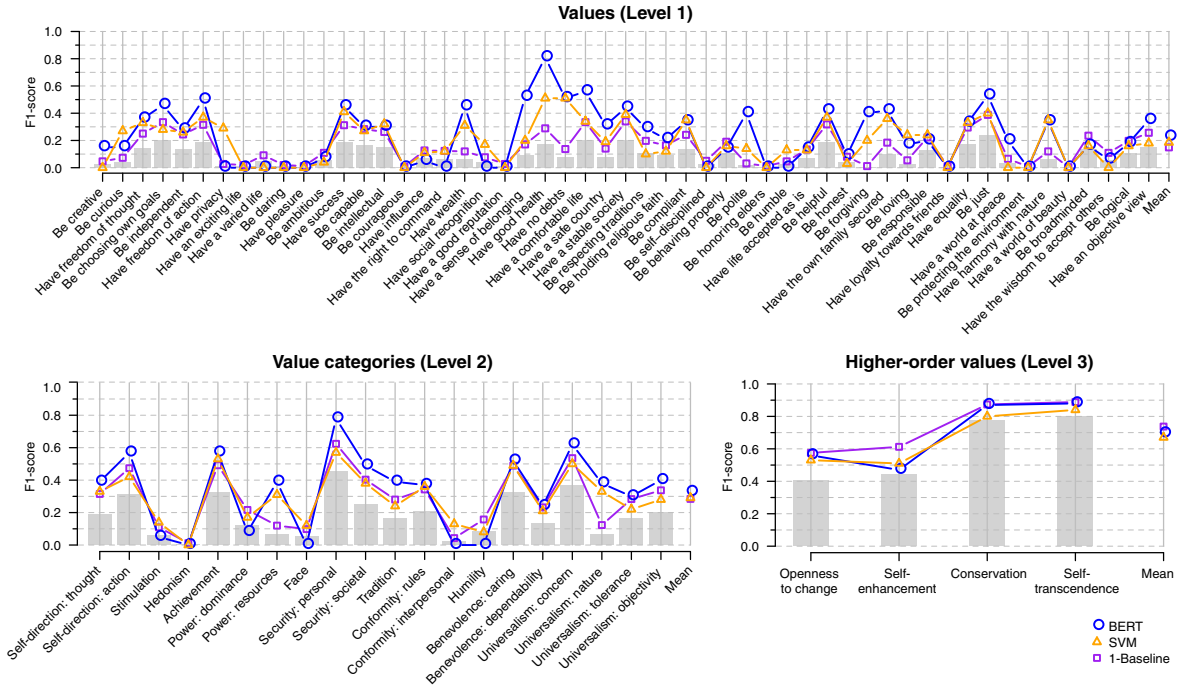

Figure 3: Parallel coordinates plot of F_1-scores on the USA test set over the labels by level. The grey bars show the label distribution, which is equal to the F_1-score of random guessing as per this distribution.

Table 4 shows the results averaged over all labels. BERT performs best according to F_1-score for Level 1 ($p = 0.007$ vs. SVM and $p = 0.001$ vs. 1-Baseline; $n = 53$) and for Level 2 ($p = 0.153$ and $p = 0.117$; $n = 20$), but is worse than or at the baseline for higher levels (n too small for test). The comparably bad performance at higher levels is somewhat surprising, as it indicates that the categories at these higher levels are harder to separate by state-of-the-art language-based approaches. Maybe hierarchical classification approaches (e.g., Babbar et al., 2013) can address this comparably weak performance by utilizing signals at each level of the hierarchy simultaneously. Moreover, while a F_1-score of 0.25 at Level 1 is encouraging for largely out-of-the-box approaches, clearly more work is needed. Though a recall of 0.19 may be acceptable for applications that not rely on completeness, a precision of 0.40 is clearly too low for practical uses.

As Figure 3 shows, however, considerably higher F_1-scores are reached by BERT for several values and value categories. Specifically, the identification works exceptionally well for the value *have good health* (F_1: 0.81) and the value-category *security: personal* (F_1: 0.78) that contains it. Other value categories with $F_1 \geq 0.5$ are *universalism: concern*, *self-direction: action*, *achievement*, and *benevolence: caring*. The out-of-the-box models thus perform reasonably well for a few selected values and categories within the USA part. Moreover, Figure 3 indicates some correlation of value frequency (grey bars) with classifier performance (colored lines). One reason for this correlation could be that the dataset is too small for training reliable classifiers on the infrequent values. Another reason might be that there is a more developed vocabulary concerning frequent values, making it easier for classifiers to identify these values. The results are distributed alongside the dataset for follow-up analyses.

Model	Level 1				Level 2				Level 3				Level 4a				Level 4b			
	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA	Afr.	Chi.	Ind.	USA
BERT	0.20	**0.21**	**0.30**	**0.25**	**0.38**	**0.37**	**0.41**	**0.34**	0.60	**0.68**	**0.71**	0.71	**0.82**	**0.88**	**0.81**	**0.92**	**0.92**	**0.91**	**0.90**	**0.96**
SVM	**0.21**	**0.21**	0.25	0.20	0.29	0.30	0.27	0.30	0.53	0.57	0.57	0.67	0.80	0.82	0.74	0.88	0.90	0.87	0.87	0.92
1-Baseline	0.16	0.13	0.12	0.16	0.27	0.23	0.21	0.28	**0.63**	0.65	0.62	**0.75**	0.80	**0.88**	0.79	**0.92**	**0.92**	**0.91**	**0.90**	**0.96**

Table 5: Macro F_1-score on each test set over all labels by level. Best scores per part and level marked bold. The scores for USA are the same as in Table 4.

5.2 Results Across Culture

For testing classification robustness, we here apply the same approaches without re-training to all test sets. The non-US parts are considerably smaller and as a result ~28% of the values are lacking arguments (cf. Table 1). However, the 1-Baseline is equally affected by this lack, thus providing for a comparison with the previous setting.

Table 5 shows the F_1-scores for each test set averaged over all labels. Once more, BERT performed best by the F_1-score for Level 1 ($p = 0.006$ vs. SVM and $p < 0.001$ vs. 1-Baseline; $n = 169$) and Level 2 (both $p < 0.001$; $n = 74$), whereas no significant difference was found for Level 3 ($p = 0.179$ and $p = 0.856$; $n = 16$). BERT and SVM perform on Level 1 and 2 similar across parts. Maybe due to the clarity of its editored arguments, BERT performs best for India, despite the 1-Baseline performing best for USA.

These findings constitute first evidence that using a cross-cultural value taxonomy could result in robust methods for identifying the values behind arguments, even though more data and research seem necessary to get there.

6 Conclusion

A computational identification of human values behind arguments is a challenging but also necessary task. With our research we contribute (1) a multi-level taxonomy with 54 values based on social science research, (2) a labeled dataset comprised of 5270 arguments from four sources, and (3) empirical analyses that cover multiple value granularity levels and compare different cultures.

Based on this work a logical next step are analyses that fully exploit relationships between labels. Hierarchical classification approaches appear promising here (e.g., Babbar et al., 2013); learning rules for multi-label classification (e.g., Loza Mencía and Jannsen, 2016) can provide insights into value-relationships.

Moreover, the dataset should be extended to include data from more cultures or territories, genres (e.g., blog posts), modalities (offline and spoken argumentation), and languages. Probably an automated translation with manual assurance, as we did for the dataset's China part, may not be sufficient. Though we optimized the annotation process, the argument acquisition requires a community effort to ensure the widest variety of data. Employing annotators from different cultures is a requirement to analyze and mitigate potential sources of bias. A subsequent step of ranking the annotated values by importance can be beneficial for certain use cases, especially when using the higher taxonomy levels.

Values are a major contributor to argument strength (Bench-Capon, 2021), and the large-scale mining from web data could improve all of argument categorization, assessment, and generation. For example, matching values between arguments could be effective for both supporting and countering arguments. Clearly expressing values behind arguments could avoid misunderstandings between humans and automated argumentation systems (Kiesel et al., 2021). Similarly, an "objective" highlighting of common values behind arguments across political camps could be a step towards resolving seemingly fundamental disagreements.

Finally, the analysis of values in large-scale text corpora can also be of interest of social science scholars. How are values expressed online? Combined with Internet archive data, one could even analyse references to values over time. We thus hope that this work can serve as a first step towards a better understanding of how the public sees and saw human values in everyday (digital) life.

7 Ethics Statement

Identifying values in argumentative texts could be used in various applications like argument faceted search, value-based argument generation, and value-based personality profiling. In all these applications, an analysis of values has the opportunity to broaden the discussion (e.g., by present-

ing a diverse set of arguments covering a wide spectrum of personal values in search or inviting people with underrepresented value-systems to discussions). At the same time, a value-based analysis could risk to exclude people or arguments based on their values. However, in other cases, for example hate speech, such an exclusion might be desirable.

While we tried to include texts from different cultures in our dataset, it is important to note that these samples are not representative of their respective culture, but intended as a benchmark for measuring classification robustness across sources. A more significant community effort is needed to collect more solid datasets from a wider variety of sources. To facilitate the inclusivity of different cultures, we adopted a personal value taxonomy that has been developed targeting universalism and tested across cultures. However, in our study, the annotations have all been carried out by annotators from a western background. Even though the value taxonomy strives for universalism, a potential risk is that an annotator from a specific culture might fail to correctly interpret the implied values in a text written by people from a different culture.

Finally, as mentioned in Section 4, we did not gather any personal information in our annotation studies, and we ensured that all our annotators get paid more than the minimum wage in the U.S.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling Frames in Argumentation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019)*, pages 2922–2932. ACL.

Milad Alshomary and Henning Wachsmuth. 2021. Toward audience-aware argument generation. *Patterns*, 2(6):100253.

Katie Atkinson and Trevor Bench-Capon. 2021. Value-based argumentation. *Journal of Applied Logics*, 8(6):1543–1588.

Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih-Reza Amini. 2013. On flat versus hierarchical classification in large-scale taxonomies. In *27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*, pages 1824–1832.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *J. Log. Comput.*, 13(3):429–448.

Trevor J. M. Bench-Capon. 2021. Audiences and argument strength. In *3rd Workshop on Argument Strength (ArgStrength 2021)*.

Duane Brown and R. Kelly Crace. 2002. Life values inventory facilitator's guide.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic. Association for Computational Linguistics.

An-Shou Cheng and Kenneth R. Fleischmann. 2010. Developing a meta-inventory of human values. In *73rd ASIS&T Annual Meeting (ASIST 2010)*, volume 47, pages 1–10. Wiley.

Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal & document design*, 13(1).

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Charlie Egan, Advaith Siddharthan, and Adam Z. Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

George W. England. 1967. Personal value systems of american managers. *Academy of Management journal*, 10(1):53–68.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, pages 390–397.

Gilad Feldman. 2021. Personal values and moral foundations: Examining relations and joint prediction of moral variables. *Social Psychological and Personality Science*, 12(5):676–686.

Roni Friedman, Lena Dankin, Yoav Katz, Yufang Hou, and Noam Slonim. 2021. Overview of KPA-2021 shared task: Key point based quantitative summarization. In *Proceedings of the 8th Workshop on Argumentation Mining*. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *34th AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 7805–7813. AAAI Press.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2020. World values survey: Round seven - country-pooled datafile.

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

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1120–1130. Association for Computational Linguistics.

Lynn R Kahle, Basil Poulos, and Ajay Sukhdial. 1988. Changes in social values in the united states during the past decade. *Journal of Advertising Research*, 28(1):35–41.

Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *3rd Conference on Conversational User Interfaces (CUI 2021)*, New York. ACM.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality & quantity*, 38:787–800.

Eneldo Loza Mencía and Frederik Jannsen. 2016. Learning rules for multi-label classification: a stacking and a separate-and-conquer approach. *Machine Learning*, 105:77–126.

Tushar Maheshwari, Aishwarya N. Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. 2017. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 731–741. Association for Computational Linguistics.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2015. Argumentation mining in parliamentary discourse. In *Principles and practice of multi-agent systems*, pages 16–25. Springer.

Milton Rokeach. 1973. *The nature of human values*. New York, Free Press.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50:19–45.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4).

John R Searle. 2003. *Rationality in action*. MIT press.

Juan Carlos Teze, Antoni Perello-Moragues, Lluís Godo, and Pablo Noriega. 2019. Practical reasoning using values: an argumentative approach based on a hierarchy of values. *Annals of Mathematics and Artificial Intelligence*, 87(3):293–319.

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.

Thomas L. van der Weide, Frank Dignum, John-Jules Ch. Meyer, Henry Prakken, and Gerard Vreeswijk. 2009. Practical reasoning using values. In *Argumentation in Multi-Agent Systems (ArgMAS 2009)*, volume 6057 of *Lecture Notes in Computer Science*, pages 79–93. Springer.

Rand R. Wilcox. 1996. *Statistics for the Social Sciences*. Academic Press Inc.

A Annotation Interface

Figure 4 and 5 show screenshots of the custom annotation interface. Its source code is distributed as part of the dataset at `https://github.com/webis-de/ACL-22`.

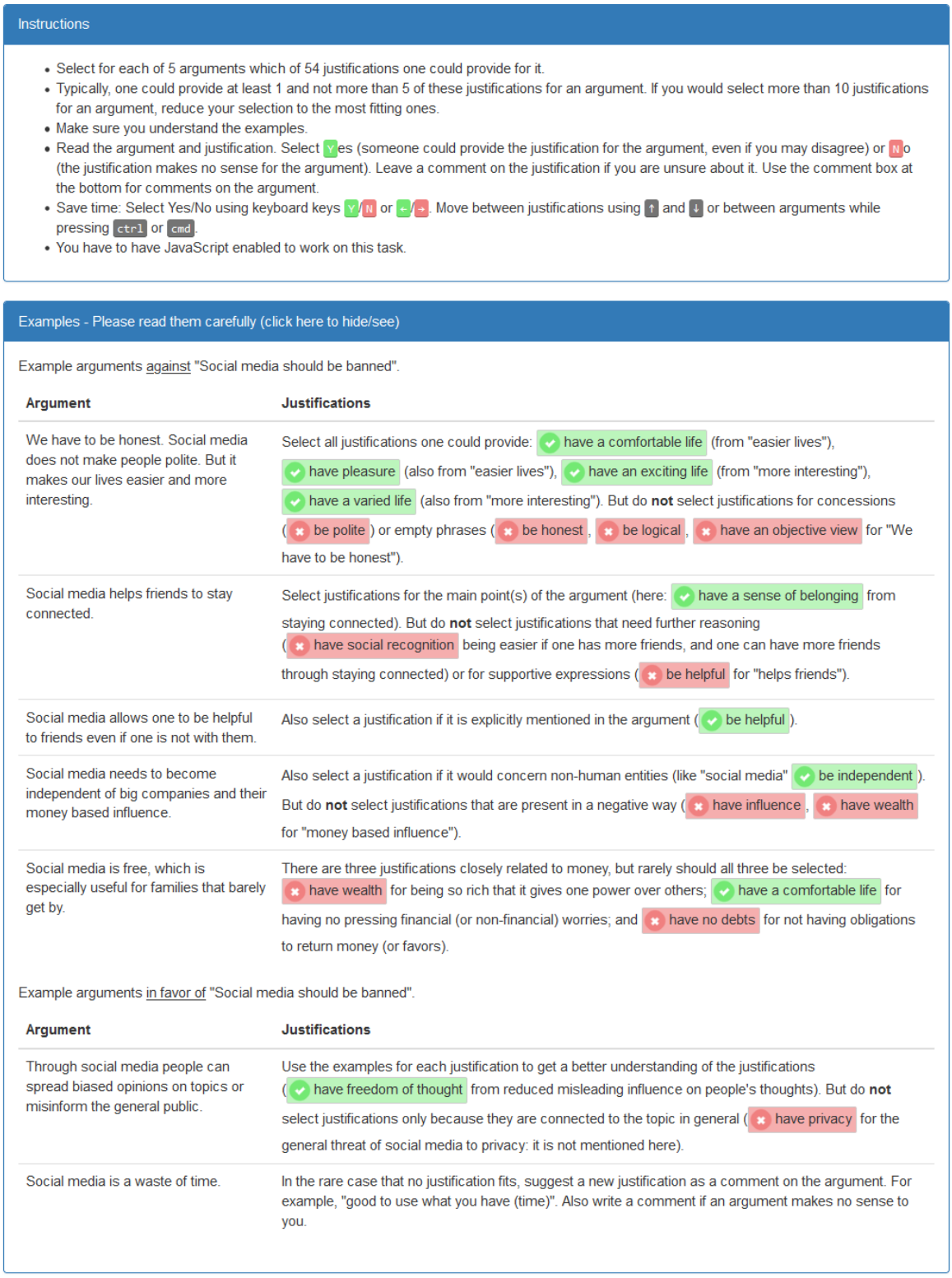

Figure 4: Screenshot ot the first part of the annotation interface, containing instructions and examples.

Figure 5: Screenshot ot the second part of the annotation interface, which consists of three panels: (1) the top left panel places the argument in a scenario ("Imagine"); (2) the top right panel formulates the annotation task for a value (here: *have wealth*) as a yes/no question, describing the value with examples; and (3) the bottom panel shows the annotation progress for the argument and allows for a quick review of selected annotations.