The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments

Nailia Mirzakhmedova,[1] **Johannes Kiesel,**[1] **Milad Alshomary,**[2] **Maximilian Heinrich,**[1] **Nicolas Handke,**[3] **Xiaoni Cai,**[4] **Valentin Barriere,**[5] **Doratossadat Dastgheib,**[6] **Omid Ghahroodi,**[7] **Mohammad Ali Sadraei Javaheri,**[7] **Ehsaneddin Asgari,**[8] **Lea Kawaletz,**[9] **Henning Wachsmuth,**[2] **Benno Stein**[1]

[1]Bauhaus-Universität Weimar, [2]Leibniz University Hannover, [3]Universität Leipzig, [4]Technische Universität München, [5]CENIA, [6]Shahid Beheshti University, [7]Sharif University of Technology, [8]Qatar Computing Research Institute, [9]Heinrich-Heine-Universität Düsseldorf

Abstract

While human values play a crucial role in making arguments persuasive, we currently lack the necessary extensive datasets to develop methods for analyzing the values underlying these arguments on a large scale. To address this gap, we present the Touché23-ValueEval dataset, an expansion of the Webis-ArgValues-22 dataset. We collected and annotated an additional 4 780 new arguments, doubling the dataset's size to 9 324 arguments. These arguments were sourced from six diverse sources, covering religious texts, community discussions, free-text arguments, newspaper editorials, and political debates. Each argument is annotated by three crowdworkers for 54 human values, following the methodology established in the original dataset. The Touché23-ValueEval dataset was utilized in the SemEval 2023 Task 4. ValueEval: Identification of Human Values behind Arguments, where an ensemble of transformer models demonstrated state-of-the-art performance. Furthermore, our experiments show that a fine-tuned large language model, Llama-2-7B, achieves comparable results.

Keywords: Corpus (Creation, Annotation, etc.); Document Classification, Text categorisation

1. Introduction

Why might one person find an argument more persuasive than someone else? One answer to this question is rooted in the values they hold and, more specifically, in the priority they give to those values. For example, should "having privacy" be considered more important than "having a safe country"? Such differences in value prioritization can prevent people from finding common ground on debatable topics or even amplify disagreements. Moreover, such disparities extend beyond individual differences, manifesting as cultural variations that further contribute to disputes. Cultural norms can strongly influence the prioritization of values, leading to distinct perspectives on a range of topics.

In computational linguistics, incorporating human values can provide context for categorizing, comparing, and evaluating argumentative statements. This approach assists various applications: facilitating social science research on values using large-scale datasets; assessing arguments, considering cultural and social perspectives; generating or selecting arguments for a target audience; and identifying opposing and shared values on both sides of a controversial topic. The most widely recognized and validated value categorization is that proposed by Schwartz et al. (2012), shown in Figure 1.

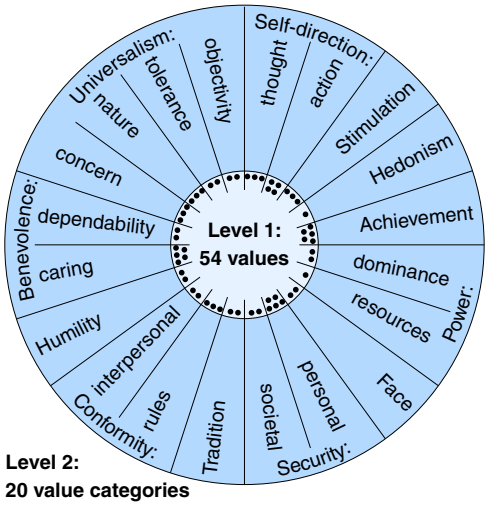

Figure 1: The employed value taxonomy of 20 value categories and their associated 54 values (shown as black dots) as defined by Kiesel et al. (2022). Categories that tend to conflict are placed on opposite sites. Illustration adapted from Schwartz et al. (2012): "universalism: objectivity" was added based on alternate categorizations.

To address the challenges of human value identification—such as the wide variety of values, their often implicit use, and their ambiguous

Contact: nailia.mirzakhmedova@uni-weimar.de

LREC-COLING 2024, pages 16121–16134

Argument source	Year	Arguments				Unique conclusions			
		Train	Validation	Test	∑	Train	Validation	Test	∑
Main dataset									
IBM-ArgQ-Rank-30kArgs	2019–20	4576	1526	1266	7368	46	15	10	71
Conf. on the Future of Europe	2021–22	591	280	227	1098	232	119	80	431
Group Discussion Ideas	2021–22	226	90	83	399	54	23	16	93
∑ (main)		5393	1896	1576	8865	332	157	106	595
Supplementary dataset									
Zhihu	2021	-	100	-	100	-	12	-	12
Nahj al-Balagha	900–1000	-	-	279	279	-	-	81	81
The New York Times	2020–21	-	-	80	80	-	-	80	80
∑ (supplementary)		-	100	359	459	-	12	161	173
∑ (complete)		5393	1996	1935	9324	332	169	267	768

Table 1: Key statistics of the main and supplementary datasets by argument source.

definition—Kiesel et al. (2022) previously developed the practical foundations for AI-based systems: a consolidated multi-level taxonomy based on extensive categorization by social scientists and an annotated dataset of 5 270 arguments, the Webis-ArgValues-22 (Kiesel et al., 2022). However, this dataset has two main shortcomings: (i) its size is relatively small for training classifiers that need to capture the (yet unknown) features of each human value; (ii) 95% of its arguments originate from a single background (the USA), restricting the development of cross-cultural value detection models.

In this work, we aim to address these limitations by introducing an expanded dataset following the same taxonomy (cf. Figure 1): the Touché23-ValueEval dataset. It includes 9 324 arguments on diverse statements collected from various sources: free-text arguments (IBM-ArgQ-Rank-30kArgs), political debates (Conference on the Future of Europe), group discussions (Group Discussion Ideas), community dialogues (Zhihu), religious texts (Nahj al-Balagha), and newspaper editorials (The New York Times). The presented expansion broadens the diversity of arguments in terms of cultures, territories, and historical perspectives (the Nahj al-Balagha dates back over 1 000 years ago). As a result, we quadruple the proportion of arguments from a non-USA background from 5% to 20%. The proposed dataset was collected and annotated for the SemEval 2023 Task 4. ValueEval: Identification of Human Values behind Arguments (Kiesel et al., 2023)[1] and is publicly available online.[2]

In the following, we detail the construction of the Touché23-ValueEval dataset (Section 3 and 4), provide a statistical overview (Section 5), and compare selected approaches on the dataset, including the 1-Baseline, BERT-based approach, LLM-based approaches, and the competition-winning ensemble-based approach (Section 6).

[1] https://touche.webis.de/semeval23/touche23-web

[2] Dataset: https://doi.org/10.5281/zenodo.6814563

2. Related Work

Human values have been the focus of an extensive line of research. Rokeach (1973) first defined values as specific end states or modes of conduct that humans desire. Rokeach then introduced the value system as a prioritization of these values based on cultural, social, and personal factors. The author developed a practical survey of 36 values that distinguishes between end states and behavior. For cross-cultural analysis, Schwartz (1994); Schwartz et al. (2012) derived 48 value questions from universal individual and societal needs, including concepts such as "obeying all the laws" and "being humble". Cheng and Fleischmann (2010) consolidated 12 taxonomies into a "meta-inventory" with 16 values, demonstrating considerable overlap.

In certain computational frameworks of argumentation, the strength of an argument depends on the audience's preferences which are shaped by their values. For example, 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).To operationalize these frameworks, it is crucial to automatically identify human values in arguments. Pioneering work in this area was done by Kiesel et al. (2022), who presented the Webis-ArgValues-22 (Kiesel et al., 2022) dataset, which we build upon in this work.

Outside of argumentation, several works in natural language processing utilize values. For example, in the context of interactive systems, Ammanabrolu et al. (2022) aim to tune interactive chat-based agents towards morally acceptable behavior. However, since their operationalization of values is limited to valence (good or bad) and target (self or other), the model can not explain in abstract terms why something would be acceptable. Liu et al. (2023) follow a similar approach based on human edits that change the text to morally acceptable ("value-aligned") behavior. A related dataset to the

ours is ValueNet by Qiu et al. (2022),[3] which contains 21K one-sentence descriptions of social scenarios (taken from SOCIAL-CHEM-101 of Forbes et al. (2020)) annotated for the 10 value categories from an earlier version of Schwartz's value taxonomy. A major difference to our dataset is the more ordinary situations in ValueNet (e.g., whether to say "I miss mom"). Unlike ValueNet, where scenario descriptions could be seen as conclusions and "utility" annotations (-1 to +1) as stances, our dataset explicitly targets the underlying premise that links value category and description, highlighting the premise's role as the source of ethical reasoning.

3. Collecting Arguments

To explore approaches for automated human value detection in arguments, we collected a dataset of 9 324 arguments. Aligned with the Webis-ArgValues-22 dataset (Kiesel et al., 2022), 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. Notably, our dataset includes a significant portion (4 755; 51%) of novel arguments. While some of these (3 298; 69%) were derived from the same sources as Webis-ArgValues-22, the remaining arguments (1 457; 31%) were obtained from three entirely new sources.

The dataset is split for usage in classification (train, validation, and test sets). It is also further categorized into a rather cohesive "main" dataset of 8 865 arguments and a more diverse (in terms of both written form and ethical reasoning) "supplementary" dataset of just 459 arguments. Table 1 provides key figures for the data. The main dataset contains arguments from three sources with similar value distributions, and its arguments are split randomly into train, validation, and test. To avoid train-test leakage from argument similarity, we ensured that all arguments with the same conclusions (but different premises) were in the same set. Whereas the main dataset represents a traditional in-domain classification setup, the supplementary dataset is intended for evaluating the robustness of classifiers on diverse data. Therefore, it should be used only for validation and testing.

The following sections describe each data source, the collection process, and the preprocessing of the arguments. For illustration, Table 2 provides one example argument per source.

3.1. IBM-ArgQ-Rank-30kArgs

The original Webis-ArgValues-22 dataset contains 5 020 arguments from the IBM-ArgQ-Rank-30kArgs dataset (Gretz et al., 2020). We add 2 999 more arguments from this source. However, to avoid train-test leakage as mentioned previously, we excluded 651 arguments of the Webis-ArgValues-22 for which the conclusion is in the new test set.

Source The authors tasked crowdworkers to write one supporting and one contesting argument on one of 71 controversial topics. The dataset totals 30 497 arguments, each of which is rated for quality by crowdworkers, with the "high-quality" criterion being "if a person preparing a speech on the topic will be likely to use the argument as is in [their] speech." (Gretz et al., 2020)

Collection process Similar to the methodology employed in Webis-ArgValues-22, we sampled premises that were designated as "high-quality" by at least 50% of the crowdworkers. The corresponding topics associated with these premises were then utilized as conclusions.

Preprocessing We also followed the same preprocessing approach: we manually corrected encoding errors in the text body of each argument, ensured a uniform character set for punctuation, and formatted arguments to be HTML compatible.

3.2. Conference on the Future of Europe

Our dataset incorporates 1 098 arguments that are pro or con 431 unique conclusions, collected from the Conference on the Future of Europe portal.[4]

Source Conference on the Future of Europe (CoFE) was an online participatory democracy platform intended to involve citizens, experts, and EU institutions in a dialogue focused on the future direction and legitimacy of Europe. CoFE was designed as a user-led series of debates, where anyone could submit a proposal in any of the EU24 languages. For each proposal, any other user could endorse or criticize the proposals (akin to the functionality of a "like" on social media), facilitating open commentary, discussions, and responses to other users' viewpoints.

Collection Process We used the CoFE dataset (Barriere et al., 2022; Barriere and Balahur, 2023), which includes more than 20 000 comments on around 4 200 proposals in 26 languages (primarily English, German, and French). About 35% of the comments in the dataset were labeled by users themselves to indicate their opinion on the proposal, roughly 6% were annotated by experts, while the rest remain unlabeled.

[3] Not related to ValueNet by Giorgis et al. (2022).

[4] https://futureu.europa.eu

Argument	Value categories	Source
○ Con “We should end the use of economic sanctions”: Economic sanctions provide security and ensure that citizens are treated fairly.	Security: societal, Universalism: concern	IBM-ArgQ-Rank-30kArgs
○ Pro “We need a better migration policy.”: Discussing what happened in the past between Africa and Europe is useless. All slaves and their owners died a long time ago. You cannot blame the grandchildren.	Universalism: concern	Conf. on the Future of Europe
○ 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.	Security: societal, Universalism: concern	Group Discussion Ideas
○ Con “We should secretly give our help to the poor”: By showing others how to help the poor, we spread this work in the society.	Benevolence: caring, Universalism: concern	Nahj al-Balagha
○ Con “We should crack down on unreasonably high incomes.”: If the key to an individual’s standard of living does not lie in income, then it is useless to simply regulate income.	Security: personal, Universalism: concern	Zhihu
○ Pro “All of this is a sharp departure from a long history of judicial solicitude toward state powers during epidemics.”: In the past, when epidemics have threatened white Americans and those with political clout, courts found ways to uphold broad state powers.	Power: dominance, Universalism: concern	The New York Times

Table 2: Six example arguments (stance, conclusion, and premise) and their annotated value categories. We selected these to showcase different ways for resorting to the value “be just”, which belongs to the category “Universalism: concern”.

Preprocessing Due to time constraints, we use only proposals written in English. Out of 6 985 user-annotated comment-proposal pairs in the CoFE dataset, we preprocessed 1 098 comments from 431 discussions. We manually identified a conclusion in each proposal and one or more premises in the corresponding comments. We manually ensured that the arguments had a similar length and structure to those in Webis-ArgValues-22.

3.3. Group Discussion Ideas

We extended the 100 arguments of the “India” part of the Webis-ArgValues-22, collected from the Group Discussion Ideas web page[5] by including 299 new arguments from the same source.

Source This website collects (in English) pros and cons on various topics covered in Indian news “to provide all the valid points for the trending topics, so that the readers will be equipped with the required knowledge” for a group discussion or debate. The web page currently lists a team of 16 authors. We got their permission to distribute the arguments.

Collection process We crawled the web page and semi-automatically extracted arguments. The original 100 arguments were from the 2021 section of the web page. We collected the additional 299 arguments from all topics from 2022.

Preprocessing We manually ensured that the arguments had a similar structure to those in the Webis-ArgValues-22 dataset by rewording and slightly shortening them if necessary.

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

3.4. Zhihu

The original Webis-ArgValues-22 dataset contains 100 arguments from Zhihu as its “China” part. We incorporated all of these into the Touché23-ValueEval dataset as is.

Source Zhihu is a Chinese question-answering website.[6] Arguments were taken from the recommendation and hotlist section.

Collection process Kiesel et al. (2022) manually identified and rephrased key points (premises and conclusions) in the answers.

Preprocessing Kiesel et al. (2022) automatically translated the key points into English and corrected the translation manually.

3.5. Nahj al-Balagha

We obtained 279 arguments from the Nahj al-Balagha, a collection of Islamic religious texts, featuring advice and arguments on moral behavior.

Source The book Nahj al-Balagha contains moral aphorisms and eloquent speeches attributed to Ali ibn Abi Talib (600 CE, though published centuries later), who is known as one of the main Islamic elders. The Nahj al-Balagha includes more than 200 sermons, 80 letters, and 500 sayings. Originally written in Arabic, Nahj al-Balagha has been translated into various languages. We employed the Farsi translations of the book.

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

Collection process We first manually extracted 302 premises from the Nahj al-Balagha: 181 were extracted verbatim and 121 were distilled from the text. We manually derived conclusions, consolidating similar ones. To maintain a balanced perspective, we rephrased a few statements to oppose the conclusions. All 279 arguments used in our annotations are from this pool of 302 statements, with 23 unclear arguments excluded from annotation.

Preprocessing Preliminary translations were carried out from Farsi into English using automated translation, which were subsequently reviewed by the native speakers of Farsi.

3.6. The New York Times

We collected 80 arguments from editorials published in The New York Times.[7] Since the arguments are restricted by copyright, we provide software to extract the arguments from the Internet Archive in a reproducible manner.[8]

Source The New York Times is a renowned US-American daily newspaper that is available in print and via an online subscription.

Collection process The arguments were extracted from 12 editorials, published between July 2020 and May 2021, tagged with at least one of "coronavirus (2019-ncov)", "vaccination and immunization", or "epidemics". We selected articles with an overall high quality of argumentation, as assessed by three linguistically trained annotators.

Preprocessing The premises, conclusions, and stances were manually identified by four annotators (three per text), and curated by two linguist experts. The curators selected 80 especially clear arguments for the dataset.

4. Crowdsourcing Value Annotations

For consistency, we replicated the crowdsourcing setup that was used for Webis-ArgValues-22 (see Kiesel et al., 2022 for details). For illustration, we reprint the screenshots of the annotation interface in the Appendix. As the screenshots show, the interface contains annotation instructions (cf. Figure 7) and uses yes/no questions for labeling each argument for each of the 54 level 1 values (cf. Figure 8). We were able to recruit 13 of the 27 original annotators, who were offered the same payment.

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

[8] https://github.com/touche-webis-de/touche-code/tree/main/semeval23/human-value-detection/nyt-downloader

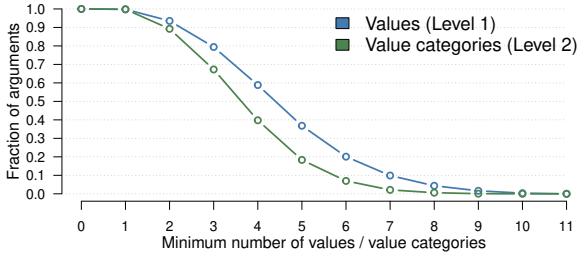

Figure 2: Fraction of arguments in the complete dataset having a minimum number of assigned values (out of 54) or value categories (out of 20).

In total, the annotators made 774 360 yes/no annotations for 4 780 new arguments. The overall inter-annotator agreement (as measured by Krippendorf's alpha) is $\alpha = 0.41$.

We adopted MACE (Hovy et al., 2013) to fuse the annotations into a single ground truth. For quality assurance, we curated all annotations for arguments from complicated sources (Nahj al-Balagha and the New York Times) and for which MACE reported a low confidence in the fusion. For the curation process, we had one expert on human values (one of the authors) manually review 727 arguments for which crowdworkers disagreed.

5. Analyzing the Dataset

This section provides a quantitative overview of the collected dataset using text and value statistics.

Text statistics The dataset comprises 9 324 unique premise-conclusion pairs. As Figure 2 shows, 94% of the arguments have at least 2 values assigned to them, and 89% have more than 2 value categories. A total of 18 arguments (~0.19%) have no assigned values (i.e., they resort to no ethical judgment).[9] As seen in Table 3, premises have an average length of 23.53 words, while conclusions are shorter at an average length of 6.48 words. The stance distribution is generally balanced, with a slight skew of approximately 10% towards "pro".

Value statistics Figure 3(a) illustrates the distribution of value categories for the main dataset. All three sources exhibit similar patterns, with slight variations. Notably, arguments originating from discussion boards (Group Discussion Ideas, Conference on the Future of Europe) tend to prioritize "Universalism: Objectivity" more than those from IBM-ArgQ-Rank-30kArgs. Across all sources, the most prevalent category is "Universalism: Concern", with the least common being "Hedonism" and "Humility".

[9] The complete list of arguments with no assigned values can be found in Table 7 in the Appendix.

Argument source	Mean length		Arguments	
	Concl.	Premise	Pro	Con
IBM-ArgQ-Rank-30kArgs	5.55	19.84	3824	3544
Conf. on the Future of Europe	11.35	39.59	750	348
Group Discussion Ideas	7.87	45.27	250	149
Zhihu	8.19	27.51	59	41
Nahj al-Balagha	5.58	22.40	224	55
The New York Times	20.20	22.87	69	11
$\sum$ (complete)	6.48	23.53	5176	4148

Table 3: Statistics of mean length in words (space-separated tokens) of conclusions/premises and number of arguments per stance by source.

For the supplementary dataset, Figure 3 (b) reveals more distinct value category distributions, reflecting the difference in genre and moral reasoning between the argument sources. Arguments collected from Zhihu exhibit a high frequency of "Achievement" and "Security: Societal" value categories. For the New York Times part, "Security: Personal" is by far the most frequent value category—being behind more than 30% of the arguments—, likely influenced by its coverage of the pandemic. In contrast, the arguments from Nahj al-Balagha generally resort to fewer values, with smaller peaks for "Achievement" and "Security: Personal".

Table 4 details the distribution of (level 1) values for each source and the entire dataset. For instance, the table illustrates the emphasis on the "Have good health" value in the New York Times (NYT) part and highlights a strong contrast between "Have equality" and "Be just" in the Conference on the Future of Europe (CoFE) part.

6. Baseline Experiments

In this section, we present the results of baseline experiments for the automatic identification of human values behind arguments using the extended dataset, and we compare them with the results presented in the initial work of Kiesel et al. (2021). This task is a multi-label classification, i.e., zero or more labels are assigned to each argument.

6.1. Approaches

For both tasks of value detection (level 1) and value categories detection (level 2), we used the following approaches:

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

BERT We used the code of Kiesel et al. (2021) to fine-tune the bert-base-uncased model on the main train set of Touché23-ValueEval. We used a batch size of 8 and a learning rate of $2 \cdot 10^{-5}$ (20 epochs).

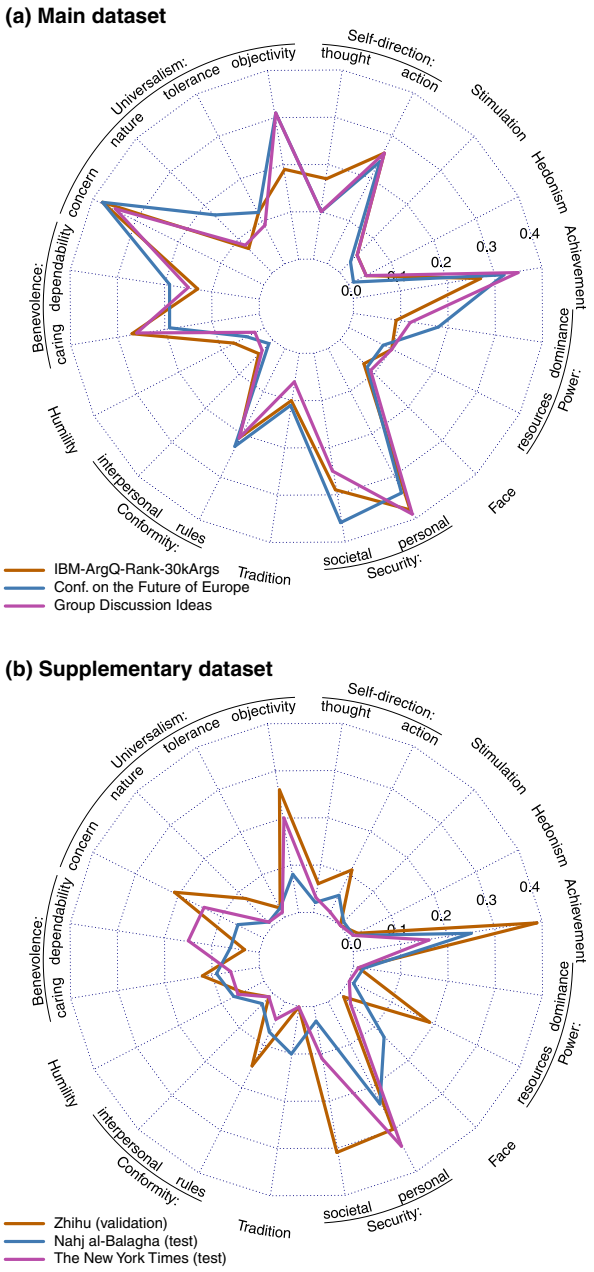

Figure 3: Distribution of value categories (level 2) across the argument sources per dataset.

Llama-2-7B The out-of-the-box Llama 2 model with seven billion parameters (Touvron et al., 2023), zero-shot prompted to answer whether an argument resorts to a specific value or value category (majority over five runs). Figure 4 illustrates the employed prompt templates.

Llama-2-7B-FT The Llama-2-7B model fine-tuned for each value and value category on the main train set (same prompt template and majority voting). We use low-rank adaptation (Hu et al., 2022) with update $r = 64$ and scaling $\alpha = 16$ to all linear layers and fine-tune for one epoch with a batch size of 8 and a learning rate of $2 \cdot 10^{-4}$.

Level		Dataset frequency (size; cf. Section 3)						
2) Value category	1) Value	IBM (7368)	CoFE (1098)	GDI (399)	Zhihu (100)	Nahj (279)	NYT (80)	Total (9324)
Self-direction: thought	Be creative	0.026	0.025	0.018	0.040	0.004	0.000	0.025
	Be curious	0.045	0.027	0.045	0.030	0.004	0.025	0.041
	Have freedom of thought	0.117	0.054	0.045	0.000	0.014	0.000	0.101
Self-direction: action	Be choosing own goals	0.129	0.105	0.103	0.030	0.004	0.000	0.119
	Be independent	0.102	0.109	0.098	0.030	0.011	0.000	0.098
	Have freedom of action	0.181	0.120	0.098	0.030	0.029	0.000	0.163
	Have privacy	0.017	0.012	0.063	0.040	0.004	0.012	0.018
Stimulation	Have an exciting life	0.017	0.004	0.018	0.000	0.000	0.000	0.015
	Have a varied life	0.038	0.027	0.040	0.000	0.004	0.000	0.035
	Be daring	0.010	0.007	0.000	0.000	0.004	0.000	0.009
Hedonism	Have pleasure	0.038	0.005	0.040	0.020	0.014	0.012	0.033
Achievement	Be ambitious	0.042	0.046	0.068	0.050	0.047	0.000	0.043
	Have success	0.120	0.097	0.148	0.160	0.068	0.012	0.116
	Be capable	0.159	0.215	0.253	0.200	0.068	0.100	0.167
	Be intellectual	0.067	0.040	0.080	0.130	0.097	0.062	0.066
	Be courageous	0.010	0.008	0.003	0.000	0.022	0.012	0.009
Power: dominance	Have influence	0.057	0.101	0.088	0.010	0.011	0.000	0.061
	Have the right to command	0.037	0.100	0.045	0.000	0.007	0.012	0.043
Power: resources	Have wealth	0.099	0.084	0.100	0.190	0.014	0.000	0.095
Face	Have social recognition	0.047	0.055	0.068	0.000	0.032	0.000	0.048
	Have a good reputation	0.022	0.040	0.028	0.010	0.111	0.025	0.027
Security: personal	Have a sense of belonging	0.077	0.108	0.075	0.010	0.075	0.025	0.080
	Have good health	0.136	0.066	0.125	0.030	0.036	0.275	0.124
	Have no debts	0.056	0.061	0.068	0.020	0.004	0.000	0.055
	Be neat and tidy	0.003	0.006	0.003	0.000	0.004	0.000	0.003
	Have a comfortable life	0.185	0.158	0.251	0.260	0.129	0.075	0.183
Security: societal	Have a safe country	0.185	0.226	0.160	0.030	0.007	0.062	0.180
	Have a stable society	0.190	0.237	0.135	0.300	0.029	0.075	0.189
Tradition	Be respecting traditions	0.077	0.105	0.040	0.000	0.000	0.000	0.075
	Be holding religious faith	0.046	0.008	0.023	0.000	0.100	0.000	0.041
Conformity: rules	Be compliant	0.124	0.179	0.120	0.070	0.022	0.000	0.126
	Be self-disciplined	0.028	0.016	0.020	0.030	0.025	0.012	0.026
	Be behaving properly	0.125	0.061	0.095	0.070	0.043	0.038	0.113
Conformity: interpersonal	Be polite	0.031	0.009	0.023	0.010	0.029	0.000	0.027
	Be honoring elders	0.010	0.003	0.010	0.000	0.004	0.012	0.009
Humility	Be humble	0.012	0.010	0.005	0.020	0.043	0.038	0.013
	Have life accepted as is	0.066	0.031	0.018	0.040	0.036	0.025	0.058
Benevolence: caring	Be helpful	0.139	0.122	0.133	0.030	0.039	0.038	0.132
	Be honest	0.043	0.046	0.060	0.010	0.014	0.012	0.043
	Be forgiving	0.018	0.005	0.005	0.000	0.007	0.000	0.015
	Have the own family secured	0.074	0.030	0.038	0.090	0.004	0.000	0.065
	Be loving	0.045	0.010	0.060	0.020	0.032	0.012	0.041
Benevolence: dependability	Be responsible	0.128	0.189	0.143	0.030	0.047	0.150	0.132
	Have loyalty towards friends	0.004	0.002	0.008	0.000	0.018	0.000	0.004
Universalism: concern	Have equality	0.168	0.019	0.216	0.090	0.011	0.088	0.167
	Be just	0.252	0.232	0.221	0.180	0.025	0.100	0.240
	Have a world at peace	0.077	0.084	0.030	0.000	0.029	0.012	0.073
Universalism: nature	Be protecting the environment	0.036	0.156	0.055	0.080	0.000	0.000	0.050
	Have harmony with nature	0.052	0.099	0.065	0.050	0.004	0.012	0.057
	Have a world of beauty	0.012	0.005	0.000	0.000	0.004	0.000	0.010
Universalism: tolerance	Be broadminded	0.094	0.069	0.080	0.010	0.014	0.012	0.086
	Have wisdom to accept others	0.053	0.069	0.033	0.010	0.007	0.000	0.052
Universalism: objectivity	Be logical	0.101	0.210	0.193	0.120	0.011	0.125	0.115
	Have an objective view	0.127	0.172	0.163	0.160	0.065	0.150	0.133

Table 4: The 54 values of the taxonomy and dataset frequency per source: IBM-ArgQ-Rank-30kArgs (IBM), Conference on the Future of Europe (CoFE), Group Discussion Ideas (GDI), Zhihu, Nahj al-Balagha (Nahj), and The New York Times (NYT), as well as overall dataset frequency.

Model	Values (Level 1)								Value categories (Level 2)							
	Webis-ArgValues-22				Touché23-ValueEval				Webis-ArgValues-22				Touché23-ValueEval			
	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc
1-Baseline	0.08	**1.00**	0.16	0.08	0.07	**1.00**	0.13	0.07	0.16	**1.00**	0.27	0.16	0.15	**1.00**	0.26	0.15
BERT	**0.36**	0.15	0.19	**0.93**	**0.43**	0.19	0.26	**0.94**	**0.46**	0.28	0.33	**0.86**	**0.59**	0.33	0.42	**0.88**
Llama-2-7B	0.07	0.17	0.09	0.78	0.08	0.22	0.10	0.75	0.16	0.25	0.17	0.68	0.16	0.24	0.19	0.69
Llama-2-7B-FT	0.24	0.57	**0.33**	0.85	0.38	0.41	**0.38**	0.92	0.30	0.66	**0.41**	0.75	0.50	0.55	**0.53**	0.87

Table 5: Comparison of macro precision (P), recall (R), F_1-score (F_1), and accuracy (Acc) on respective test sets of Webis-ArgValues-22 and Touché23-ValueEval (the main test set) by level. Each score is an average over scores for each value or value category.

Model	Validation								Test											
	Main				Zhihu				Main				Nahj al-Balagha				New York Times			
	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
1-Baseline	0.17	**1.00**	0.29	0.17	0.13	**1.00**	0.23	0.13	0.15	**1.00**	0.26	0.15	0.07	**1.00**	0.13	0.07	0.08	**1.00**	0.15	0.08
BERT	0.55	0.32	0.41	0.86	0.39	0.39	0.39	**0.86**	**0.59**	0.33	0.42	**0.88**	**0.28**	0.28	0.28	**0.88**	**0.20**	0.29	0.24	**0.87**
Llama-2-7B	0.18	0.27	0.22	0.67	0.12	0.30	0.17	0.68	0.16	0.24	0.19	0.69	0.07	0.22	0.11	0.72	0.06	0.13	0.08	0.70
Llama-2-7B-FT	0.52	0.50	0.51	0.86	0.35	0.55	0.43	0.85	0.50	0.55	0.53	0.87	0.27	0.48	**0.35**	0.87	0.19	0.51	**0.28**	0.83
Ensemble	**0.67**	0.75	**0.71**	**0.91**	**0.41**	0.69	**0.51**	0.84	0.51	0.62	**0.56**	0.87	0.23	0.62	0.34	0.81	0.16	0.57	0.25	0.79

Table 6: Comparison of macro precision (P), recall (R), F_1-score (F_1), and accuracy (Acc) on respective validation and test sets of Touché23-ValueEval for the task of value categories detection (level 2).

Ensemble We include the winning submission of SemEval-2023 Task 4: An ensemble of 12 transformer models (DeBERTa and RoBERTa with different optimization and training data).[10]

6.2. Results

Table 5 details the performance of the employed approaches for both level 1 and level 2. To assess the impact of dataset expansion, we report the results both for the test set of Webis-ArgValues-22 and the main test set of Touché23-ValueEval. The results reveal that while the effectiveness of the 1-Baseline slightly declined compared to Webis-ArgValues-22, the F_1-score of all other approaches increased. This suggests that, although the classification difficulty marginally increased (as evident from the label distribution) the larger dataset facilitates the training of more robust models, effectively compensating for the increased difficulty.

Table 6 presents the results of our baseline experiments for the value categories detection task (level 2) on the validation and test sets of the Touché23-ValueEval dataset. The out-of-the-box (zero-shot) Llama-2-7B demonstrates significantly lower performance when compared to its fine-tuned counterpart. This result underlines the importance of training datasets for enhancing model effectiveness. Moreover, fine-tuned Llama-7B shows a substantial improvement in F_1-score over BERT, showcasing the recent advancements in NLP. Furthermore, the fine-tuned Llama-2-7B performs on par with the competition-winning ensemble.

[10]Team "Adam Smith" (Schroter et al., 2023)

Figure 5 shows the F_1-score achieved by each approach on the three test sets. For the main test set (Figure 5 a), the fine-tuned Llama-2-7B and the Ensemble achieve a comparable performance for most values. The largest difference exists for predicting "Conformity: interpersonal" (0.45 vs. 0.65). However, for the Nahj al-Balagha part (Figure 5 b), the fine-tuned Llama-2-7B performs noticeably better for "Power: resources" (0.57 vs. 0.20; 39 arguments), but much worse for "Stimulation" (0.00 vs. 0.33; 11 arguments). The New York Times part (Figure 5 c) stands out due to its frequent use of "Security: personal," with both approaches showing comparable performance for this category. Overall, the fine-tuned Llama-2-7B and the ensemble approach perform similarly for most value categories and sources of arguments.

7. Conclusion

We presented the Touché23-ValueEval dataset for Identifying Human Values behind Arguments, comprising 9 324 arguments manually labeled for 54 values and 20 value categories. We detailed its construction and its complementary nature to the Webis-ArgValues-22 dataset, which it extends in terms of argument count, cultural variety, and historical perspective. Moreover, we reported results that suggest that the expanded dataset allows for training classifiers with better classification performance. We hope this dataset will serve as a catalyst for the development of more sophisticated and nuanced approaches for the detection of human values behind arguments, no matter their source.

```
Conclusion: <conclusion>
Premise: <premise>
Question: Does the given premise resort
to the human value "<value>"?
Answer [yes/no]:
```

Prompt template for values (level 1)

```
Description: "<value category>" implies
"<value category description>".
Conclusion: <conclusion>
Stance: <stance>
Premise: <premise>
Question: Does the given premise resort
to "<value category>"?
Answer [yes/no]:
```

Prompt template for value categories (level 2)

Figure 4: Prompt templates for the Llama-2-7B model. Placeholders with purple background are replaced with the respective part of the argument, and placeholders with teal background are replaced with the names or descriptions of the respective value or value category. The descriptions are distributed alongside the dataset.

8. Ethics Statement

Although with this work we diversify the sources of arguments compared to the work of Kiesel et al. (2022), our dataset is no representative sample of all human argumentation. Despite attempts to include diverse texts, they may not fully represent cultural nuances and serve as a benchmark for measuring classification robustness across sources. To enable cultural inclusivity, we used a universal values taxonomy tested across cultures over decades. However, annotations were done by Western annotators, introducing a potential risk of misinterpreting implied values in texts from different cultures.

Values in argumentative texts can be used in applications such as argument faceted search, argument generation, and personality profiling. Analyzing values can broaden discussions by presenting diverse arguments and inviting underrepresented views. At the same time, a value-based analysis could risk excluding people or arguments based on their values. However, such an exclusion might be desirable in other cases, such as hate speech.

We gathered no personal information about our annotators and ensured that all annotators got paid more than the minimum U.S. wage.

9. Acknowledgements

Valentin Barriere was funded by National Center for Artificial Intelligence CENIA FB210017, Basal ANID. This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu).

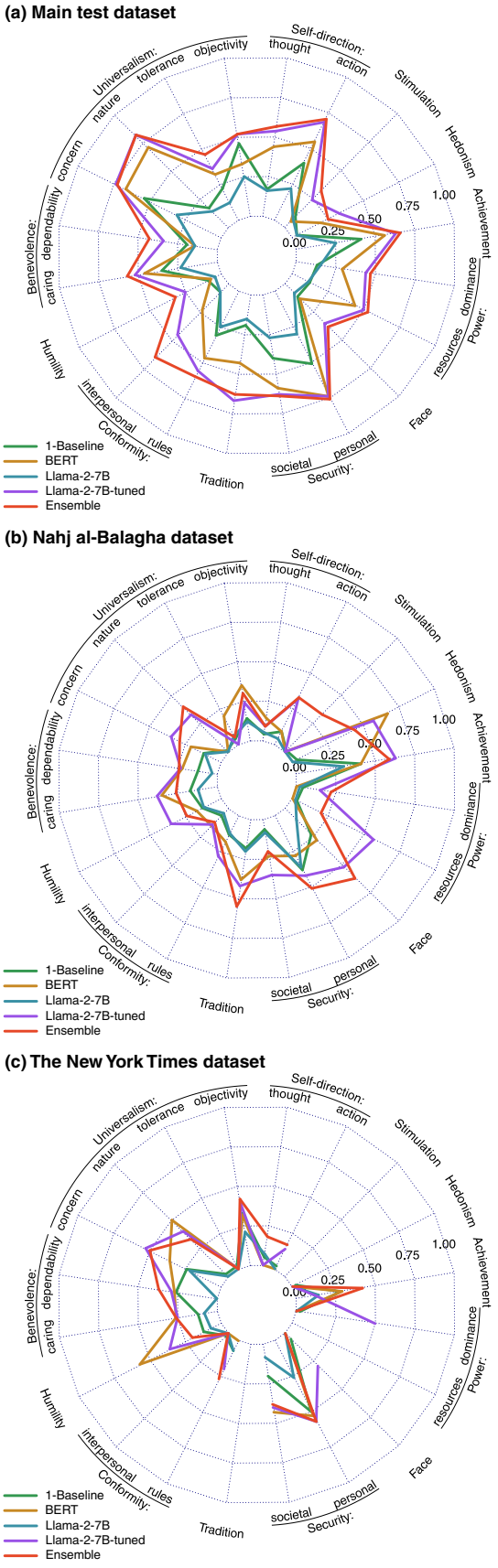

Figure 5: Detailed F_1-score for each approach per test set for the task of value categories prediction (level 2). Lines are disconnected for categories absent in the dataset (only the New York Times).

10. Bibliographical References

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to Social Norms and Values in Interactive Narratives. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'22)*, pages 5994–6017. Association for Computational Linguistics.

Valentin Barriere and Alexandra Balahur. 2023. Multilingual multi-target stance recognition in online public consultations. *Mathematics*, 11(9):2161.

Valentin Barriere, Guillaume Guillaume Jacquet, and Leo Hemamou. 2022. CoFE: A new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–422, Online only. 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.

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.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 653–670. Association for Computational Linguistics.

Stefano De Giorgis, Aldo Gangemi, and Rossana Damiano. 2022. Basic Human Values and Moral Foundations Theory in ValueNet Ontology. In *23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW'22)*, volume 13514 of *Lecture Notes in Computer Science*, pages 3–18. Springer.

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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *10th International Conference on Learning Representations (ICLR 2022)*. OpenReview.net.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

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.

Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony X. Liu, and Soroush Vosoughi. 2023. Second Thoughts are Best: Learning to Re-Align With Human Values from Text Edits. *CoRR*, abs/2301.00355.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. ValueNet: A New Dataset for Human Value Driven Dialogue System. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, pages 11183–11191. AAAI Press.

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

Daniel Schroter, Daryna Dementieva, and Georg Groh. 2023. Adam-smith at SemEval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. In *Proceedings of the 17th International Workshop*

on Semantic Evaluation (SemEval-2023), pages 532–541, Toronto, Canada. 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).

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

11. Language Resource References

Kiesel, Johannes and Alshomary, Milad and Handke, Nicolas and Cai, Xiaoni and Wachsmuth, Henning and Stein, Benno. 2022. *Webis-ArgValues-22*. Zenodo. PID https://doi.org/10.5281/zenodo.6855004.

Appendix. Figures and Tables

Figure 6 shows the label distribution to allow for a comparison with Figure 2 from Kiesel et al. (2022).

Figures 7 and 8 show screenshots of the custom annotation interface taken from Kiesel et al. (2022). Its source code is distributed as part of the Webis-ArgValues-22 dataset at `https://github.com/webis-de/ACL-22`.

Table 7 shows the arguments to which no value was assigned.

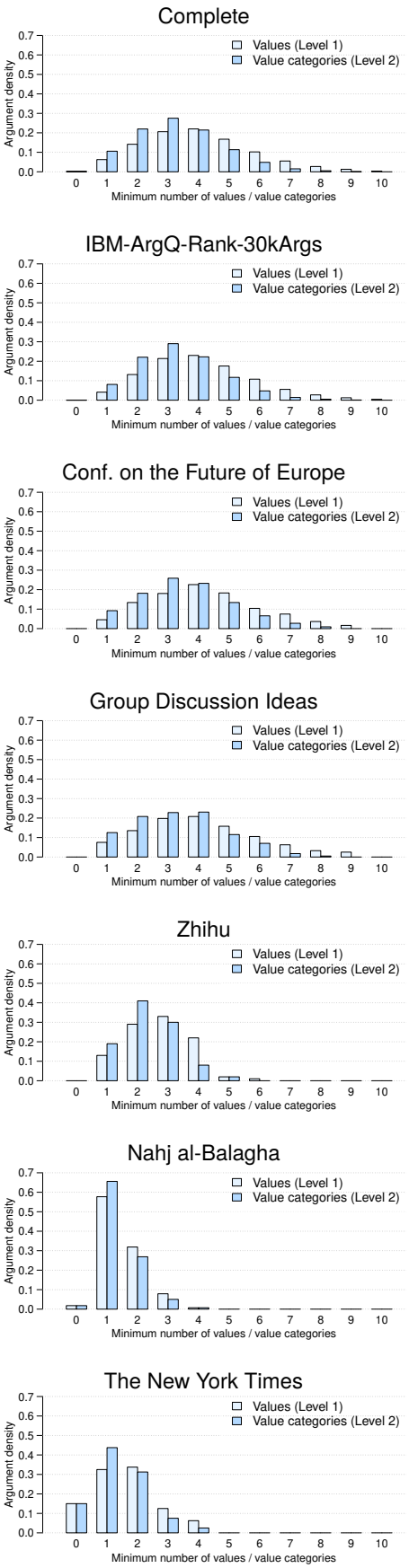

Figure 6: Fraction of arguments per dataset part having a specific number of assigned values (out of 54) or value categories (out of 10).

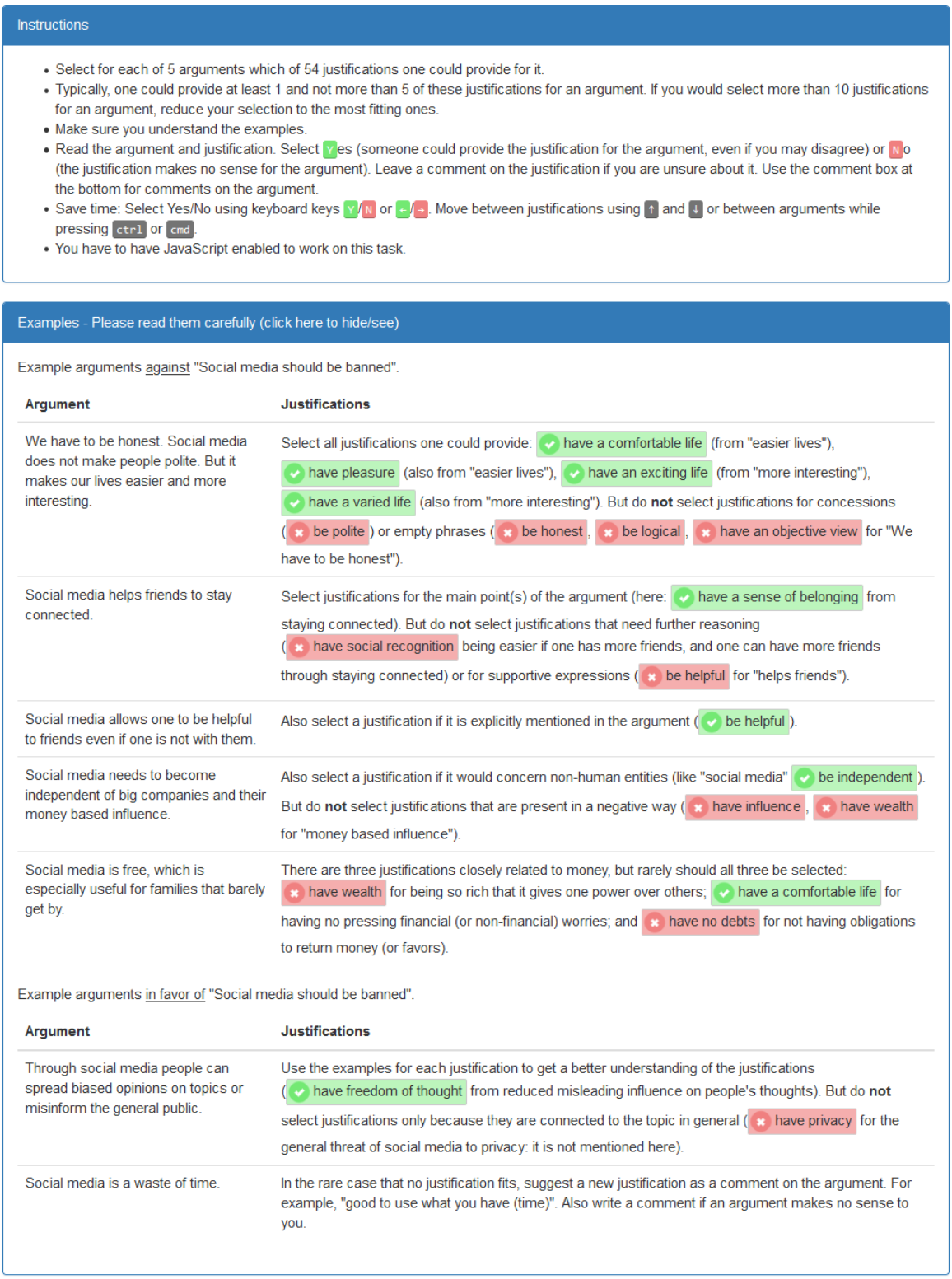

Figure 7: Screenshot of the first part of the annotation interface, containing instructions and examples.

Figure 8: Screenshot of 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.

Source/Conclusion	Premise
IBM-ArgQ-Rank-30kArgs	
◦ We should adopt a multi-party system	a multi-party system offers too many options and decreases the election voting for one candidate, spreading out the votes over too many options.
Nahj al-Balagha	
◦ Disclosing hardship humiliates humans	Disclosing hardships to those who do not have the ability to help does not have any effect but humiliating and abasing humans. In these situations, we should try not to complain and control ourselves.
◦ We should not reprove everyone who does something wrong	Every mischief monger cannot even be reproved
◦ Unintentional words and face expressions can reveals what is hidden in a man's heart	Whenever a person conceals a thing in his heart, it manifests itself through unintentional words from his tongue and (in) the expressions of his face.
◦ We should be kind to everyone	Cover your heart with kindness to people, and be friendly and kind to everyone. May you never be like a hunting animal that eats them as a trophy; Because there are two groups of people, one group is your religious brother, and the other group is like you in creation
◦ You can discover all the secrets of the world	Humans have high thinking power and the human mind is capable of understanding anything
The New York Times	
◦ Vaccination is going relatively well in this country,	although the number of people who receive a dose each day is down from its peak.
◦ Things could always get bad again, and the C.D.C. could always update its guidance and reintroduce more aggressive restrictions.	But right now, this moment feels to many like the beginning of the end of the pandemic.
◦ No other developed country is doing so badly	Graphs of the coronavirus curves in Britain, Canada, Germany and Italy look like mountains, with steep climbs up and then back down. The one for America shows a fast climb up to a plateau. For a while, the number of new cases in the U.S. was at least slowly declining. Now, according to The Times, it's up a terrifying 22 percent over the last 14 days.
◦ An epidemic that was once concentrated in blue states is increasingly raging in red ones.	Now, as New York gingerly reopens, Arizona has become a hot spot — which isn't stopping Trump from holding a rally at a Phoenix megachurch on Tuesday. Cases are also soaring in Texas, Florida and several other states.
◦ There are many knowable parameters in the equation:	your health; the prevalence of cases where you live; the safety precautions being taken any place you want to visit. But the final answer may depend on your individual risk tolerance for exposure to infectious disease.
◦ I hear too many people saying "I'm not going back to life until there's a vaccine" — as if that will immediately eliminate the risk. It won't.	Even if one of the current vaccine candidates works, it could be quite a while before it's widely distributed. And to be approved by the Food and Drug Administration, it has to protect only half of the people who take it from infection.
◦ As President Trump pushes for the quick rollout, public trust is eroding.	Only 21 percent of Americans surveyed in a CBS poll this month said they would get the vaccine as soon as possible if one was offered at no cost.
◦ It's not hard to see where this is heading: a nightmare in which we have a vaccine yet mistrust of the government is so great that people won't take it.	Three in four Democrats say if a vaccine were to become available this year, their first thought would be that it was rushed without enough testing, CBS reported.
◦ It's possible that vaccines under development by Novavax and Sanofi, which are likely to begin late-phase clinical trials later this year, may be better for the elderly, Dr. Offit noted.	Those vaccines contain immune-stimulating particles like the ones contained in the Shingrix vaccine, which is highly effective in protecting older people against shingles disease.
◦ Can it be rolled out effectively?	The Pfizer vaccine, unlike others in late-stage testing, must be kept supercooled, on dry ice around 100 degrees Fahrenheit below zero, from the time it is produced until a few days before it is injected. mRNA quickly self-destructs at higher temperatures.
◦ But it's difficult to build such road maps.	Scientists have never established correlates of immunity for pertussis, for example, although vaccines have been used against those bacteria for nearly a century.
◦ I don't blame the lucky recipients;	after all, hospitals would just offer the unused vaccine to the next person on the list.

Table 7: A complete list of arguments from the dataset with no values assigned to them.