SemEval-2023 Task 4:
ValueEval: Identification of Human Values Behind Arguments

Johannes Kiesel[1,*] **Milad Alshomary**[2] **Nailia Mirzakhmedova**[1] **Maximilian Heinrich**[1]

Nicolas Handke[3] **Henning Wachsmuth**[2] **Benno Stein**[1]

[1] Bauhaus-Universität Weimar [2] Leibniz University Hannover [3] Leipzig University

Abstract

Argumentation is ubiquitous in natural language communication, from politics and media to everyday work and private life. Many arguments derive their persuasive power from human values, such as *self-directed thought* or *tolerance*, albeit often implicitly. These values are key to understanding the semantics of arguments, as they are generally accepted as justifications for why a particular option is ethically desirable. Can automated systems uncover the values on which an argument draws? To answer this question, 39 teams submitted runs to ValueEval'23. Using a multi-sourced dataset of over 9K arguments, the systems achieved F_1-scores up to 0.87 (*nature*) and over 0.70 for three more of 20 universal value categories. However, many challenges remain, as evidenced by the low peak F_1-score of 0.39 for *stimulation*, *hedonism*, *face*, and *humility*.

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? 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, others tend to align (see Figure 1). This can cause disagreement on the best course forward, but also the support, if not formation, of political parties that promote respective highly revered values.

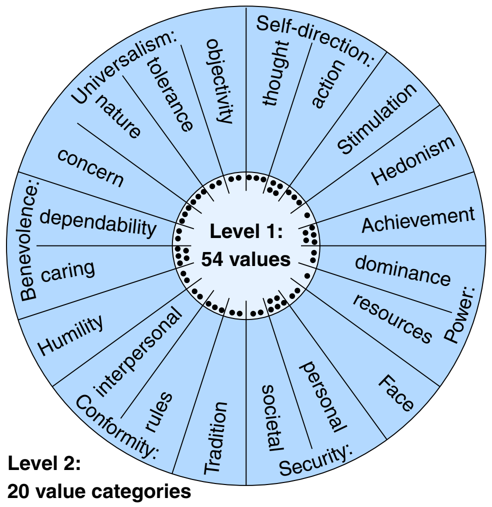


Figure 1: The employed taxonomy of 20 value categories and their associated 54 values (shown as black dots), called Level 2 and Level 1 in Kiesel et al. (2022), respectively. ValueEval'23 focused exclusively on value categories. Categories that tend to conflict are placed on opposite sites. Taxonomy and illustration are largely adapted from Schwartz (1994) and subsequent works.

Due to their outlined importance, human values have been studied both in the social sciences (Schwartz, 1994) and in formal argumentation (Bench-Capon, 2003) for decades. According to the former, 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." Consider the following example argument:

"*Social media is good for us. Though it might make people less polite, it fosters free speech.*"

To understand this argument, a reader has to acknowledge the belief (Point 1 in the definition) that the "end state" (2) of allowing for *self-directed action*, as in free speech, is desirable in general (3). To concur with the statement (4), the reader further has to prefer this state, in doubt, over *interpersonal conformity* (politeness; 5)—ignoring other arguments on the topic for the sake of the example.

The identification of human values behind arguments by means of natural language processing

[*] Contact: johannes.kiesel@uni-weimar.de

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2287–2303
July 13-14, 2023 ©2023 Association for Computational Linguistics

can assist in argument categorization, evaluation, and generation. Also, it may support the research of social scientists. The identified values can enable models of audience-specific argument strength (Bench-Capon, 2021), and it can support studies of morals (Feldman, 2021; Alshomary et al., 2022) and frames (Entman, 1993; Ajjour et al., 2019).

To advance computational methods of human value identification, we organized the ValueEval'23 shared task at SemEval as part of the Touché series.[1] Simply put, the task is to classify whether a given argument resorts to a given value (Section 3). As part of the shared task, we extended our previous dataset (Kiesel et al., 2022) to more than 9K arguments (Section 4). The task attracted submissions from 39 teams (Section 5), which we benchmarked and analyzed on our dataset (Section 6). Overall, the teams achieved respectable results, beating our BERT baseline by a large margin on all dataset parts for all 20 value categories.

2 Related Work

A dedicated line of research studied human values extensively. Rokeach (1973) first defined values as certain end states or modes of conduct that humans desire. Accordingly, they introduced the value system as a prioritization over these values based on cultural, social, and personal factors. The authors developed a practical survey of 36 values distinguishing between end states and behavior. For cross-cultural analysis, 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, revealing significant overlap. Another effort in unifying value taxonomies is the linked open data scheme ValueNet by Giorgis et al. (2022), though the authors do not compare the taxonomies as such. Based on these taxonomies are several studies in the social sciences: see Scharfbillig et al. (2021) for a recent overview and practical insights (directed at policy makers). An automated detection of human values, as is the goal of this shared task, could greatly assist such analyses (Scharfbillig et al., 2022).

Some computational frameworks of argumentation consider the strength of an argument subject to the audience's preferences defined via their values. Example frameworks 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). Automatically identifying values in natural language arguments is an important step in operationalizing these frameworks.

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 or not. Liu et al. (2023) follow a similar approach based on human edits that change text to morally acceptable ("value-aligned") behavior. A related dataset to the ours is ValueNet by Qiu et al. (2022),[2] 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 of an earlier version of Schwartz' value taxonomy. A major difference to our dataset are the more ordinary situations in ValueNet (e.g., whether to say "I miss mom"). A conceptual difference is that while ValueNet's scenario descriptions could be seen as conclusion and its "utility" annotation (-1 to +1) as stance, the link between value category and description—the premise in our data—remains implicit in ValueNet. The implicit premise is a key difference: our annotations specifically target the premise, as it is the locus of ethical reasoning.

3 Task Description

We define the task of identifying the human values behind arguments as follows:

> *Given a textual argument and a human value, classify whether the argument resorts to that value or not.*

Specifically, we employ a set of 20 value categories from our previous work (Kiesel et al., 2022) in a multi-label classification setup and evaluate approaches using macro F_1,[3] but also provided the teams with per value category F_1-scores, precision, and recall for deeper analyses. We use macro instead of micro F_1 to weigh each category the same,

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

[2] Which is not related to ValueNet by Giorgis et al. (2022).

[3] To be precise, we employ the harmonic mean of macro-averaged precision and recall.

no matter their frequency. Figure 1 shows the employed value taxonomy, which is largely based on that of Schwartz (1994) and described in full detail by Kiesel et al. (2022). For ValueEval'23, we use the 20 value categories (level 2) in a multi-label classification setup: one yes-no decision for each pair of argument and value category. We selected level 2 for the task as it is the one usually used in social science analyses (e.g., by Scharfbillig et al. (2021)), whereas level 1 is mostly used for data collection (i.e., when surveying people).

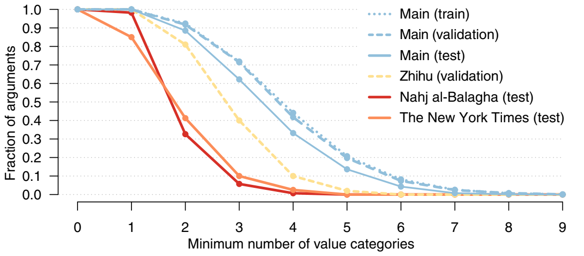


Figure 2: Fraction of arguments in each dataset split having a specific number of assigned value categories (out of 10) or more.

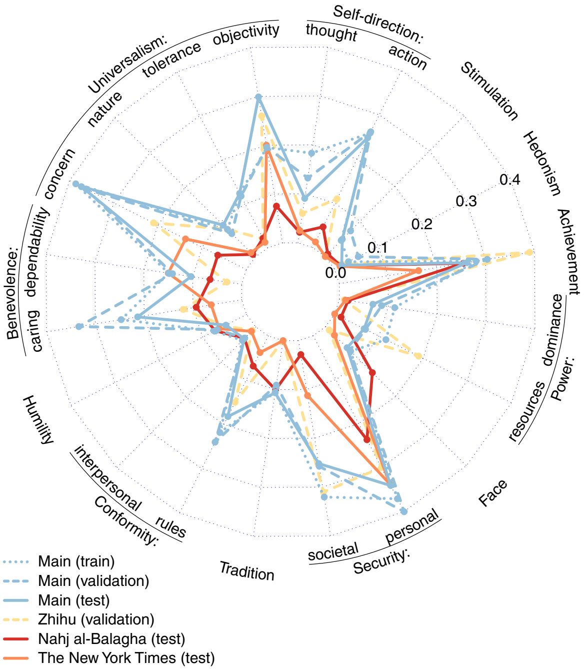


Figure 3: Ratio of arguments resorting to each value category for each dataset split.

4 Data

ValueEval'23 employs the Touché23-ValueEval dataset, which consists of 9324 arguments collected from 6 diverse sources, namely religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms. Each argument consists of three parts: two short texts—premise and conclusion—, and a stance attribute that indicates whether the premise supports the conlusion ("in favor of") or its anti-thesis ("against"). Each argument was annotated by 3 crowdworkers for 54 values, which were then mapped to 20 value categories (cf. Table 4). Mirzakhmedova et al. (2023) describe the collection and annotation in detail. The dataset is publicly available online[4] and has over 1900 downloads as of April 2023.

The dataset consists of two parts (cf. Table 2): the main dataset (8865 arguments; 95%) and the supplementary dataset (459 arguments; 5%). For the main leaderboard, we provide the main dataset as three separate sets as it is customary in machine-learning tasks: one set each for training, validation, and testing. 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. The ground truth for the test dataset has been kept secret until the camera ready deadline for ValueEval'23 participant papers.

While the main dataset reflects a classical in-domain machine learning setup, the supplementary dataset simulates an out-of-domain application. This domain difference is reflected in several aspects of the datasets. As Figure 2 shows, the majority of arguments in the main dataset have more than one (= at least two) value categories assigned to them (between 88% and 92%). As for the supplementary dataset, the Zhihu split shows a similar pattern (with 81% of arguments having more than one value category), while much fewer labels are assigned to the arguments of the New York Times (41%) and the Nahj al-Balagha split (33%). Moreover, for the latter two splits a few arguments have no value assigned (i.e., resort to no ethical judgement): 2% for Nahj al-Balagha and 15% for the New York Times. As Figure 3 shows, also the distribution of value categories is very similar within the main dataset, but quite different to the distributions for the supplementary dataset, reflecting the difference in genres and topics. Table 1 shows one example argument for each dataset source.

4.1 Main dataset

The main dataset is compiled of 8865 arguments from the following three sources, with approximately the same ratio of arguments pre source in the train, validation, and test splits (cf. Table 2):

[4] https://doi.org/10.5281/zenodo.6814563

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 1: Six example arguments (stance, conclusion, and premise) and their annotated value categories. We selected these to showcase different ways for resorting to *be just*, which is a value of the category *Universalism: concern*.

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 2: Key statistics of the main and supplementary dataset by argument source.

IBM-ArgQ-Rank-30kArgs We collected 7368 arguments from this dataset by Gretz et al. (2020), a collection of free-text arguments on 71 controversial topics that are common in the US. Arguments are collected using the US-based crowdsourcing platform Figure Eight and labeled for argumentation quality. We sampled arguments by topic, starting with that of highest quality. We used the topics as conclusions and the "arguments" as premises.

Conference on the Future of Europe We collected 1098 arguments for 431 unique conclusions from the Conference on the Future of Europe portal,[5] an online participatory democracy platform that involves citizens, experts, and EU institutions in a dialogue focused on the future direction and legitimacy of Europe. We sampled from an existing dataset (Barriere et al., 2022) proposals and corresponding comments which were originally written in English and for which the users marked their comments as supporting or contesting. We manually extracted the conclusions from the proposals and one or more premises from their comments.

Group Discussion Ideas We collected 399 arguments from the Group Discussion Ideas web page.[6] The web page aggregates pros and cons on various topics covered in Indian news to help users participate in a group discussion or debate in English.

[5] Portal: `https://futureu.europa.eu/`; dataset: `https://data.europa.eu/data/datasets/conference-on-the-future-of-europe`

[6] `https://www.groupdiscussionideas.com/`

We crawled the web-page and manually extracted one or more conclusion-premise pairs from each topic discussion, and manually labeled the stance for each pair where it was not stated explicitly.

4.2 Supplementary dataset

In addition to the main dataset, we collected a supplementary dataset of 459 arguments of different written forms and ethical reasoning. This dataset is intended as an out-of-domain challenge for submitted approaches. The arguments from the latter two sources were provided by other researchers in response to our call for data on NLP mailing lists:

Zhihu We collected 100 arguments from the recommendation and hotlist section of Zhihu, [7] a Chinese question-answering website. We translated the answers to English using automated translation and manually extracted and rephrased key points (premises and conclusions) from these.

Nahj al-Balagha We collected 279 arguments from the Nahj al-Balagha, an Islamic religious text. Conclusions and premises were deduced manually from a Farsi translation of the Arabic text, with similar conclusions being unified, and conclusions and premises translated to English. These arguments were provided by the language.ml lab.

The New York Times We collected 80 arguments from 12 news articles of The New York Times[8] that were published between July 2020 and May 2021 and contain at least one of these keywords: *coronavirus (2019-ncov)*, *vaccination and immunization*, and *epidemics*. The premises, conclusions, and stances were manually extracted by three annotators per text and curated by two linguist experts. These arguments were provided by Lea Kawaletz and Zeljko Bekcic of the Heinrich-Heine-Universität Düsseldorf.

5 Submissions

The task received submissions from 39 teams, of which 37 provided information about their approaches—including the 29 teams who submitted a paper to SemEval. The task used TIRA[9] (Fröbe et al., 2023) for evaluation. The following is a cross-sectional overview of the approaches by the 37 teams, after an overview of the top-ranked approach. For details on single approaches, refer to the papers cited in Table 3. Each team was, per test set, allowed one submission before December 16th ("early bird") and four additional submissions on January 27th. Teams were allowed more submissions after the deadline for analyses, but ground truth labels are released only after the camera-ready paper submission deadline to prevent the report of over-engineered results in the papers.

We employed anonymous submissions: teams had to choose a code name from Wikipedia's list of ethicists[10] on registration. Five teams preferred to stay anonymous and 26 other teams kept their code name as their team name.

Top-ranked Approaches Team Adam Smith uses an ensemble of 12 transformer-based models: DeBERTa and RoBERTa, both trained for either loss minimization or F_1-score maximization on three different folds each.[11] The RoBERTa models were pretrained on the full IBM-ArgQ-Rank-30KArgs dataset (Gretz et al., 2020), which is the source for most arguments in the main set (cf. Table 2). For ensembling, they averaged the predictions of the single models and used a single decision threshold that they optimized on a hold-out set. They also tried a stacked meta-classifier based on logistic regression, which performs better on the Nahj al-Balagha but not on the main test set.

Team John Arthur fine-tuned a DeBERTa model (microsoft/deberta-v2-xxlarge) on the task's data. They represented each input as a concatenation of stance, premise, and conclusion, separated by special tokens. They also found that using separate token symbols for stance ("Favour" and "Against") slightly boosted classification results. They trained their model to minimize binary cross-entropy loss. The outputs were passed through a sigmoid function to make a binary prediction for each of the value categories. A single decision threshold (0.2) was used for all categories. The team observed that having more data in the training set benefits the model's performance. Just adding as many as 100 arguments (they used the Zhihu validation set as an addition to the training set) lifted the F_1 results from 0.53 to 0.55.

Team PAI (Theodor Zwinger) used a combination of transformer models. The training data, con-

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

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

[9] https://www.tira.io/task/valueeval-at-semeval-2023-human-value-detection

[10] https://en.wikipedia.org/wiki/List_of_ethicists

[11] They report on an ablation study in their paper, which reveals that using only the three F_1-score-optimized DeBERTa models raises the F_1-score slightly.

Team		Approach				Resources			Best F_1-score		
Name (and code name if different)	Authors	Tr.	En.	NC.	Val.	🐙	🐳	🌐	Main	Nahj.	NYT
Best per value category	Several of those below								0.59	0.48	0.47
Adam Smith	Schroter et al. (2023)	✓	✓			🐙	🐳	🌐	**0.56**	**0.40**	0.27
John Arthur	Balikas (2023)	✓				🐙			0.55	-	-
PAI (Theodor Zwinger)	Ma (2023)	✓	✓			🐙			0.54	-	-
Mao Zedong	Zhang et al. (2023)	✓			✓	🐙	🐳		0.53	0.32	0.32
Confucius	Anonymous	-	-	-	-		🐳		0.53	0.31	0.31
Arthur Caplan	Song et al. (2023)	✓	✓			🐙			0.51	0.28	0.32
Hitachi (R. M. Hare)	Tsunokake et al. (2023)	✓		✓	✓				0.51	0.34	**0.34**
SUTNLP (David Gauthier)	Hematian Hemati et al. (2023)	✓			✓	🐙			0.50	0.29	-
Tübingen (Stanley Grenz)	Can (2023)	✓				🐙			0.50	0.35	-
Georg Simmel	Tian Wei et al.	✓							0.50	-	-
T. M. Scanlon	Molazadeh Oskuee et al. (2023)	✓				🐙	🐳		0.49	-	-
CSECU-DSG (Fazlur Rahman)	Aziz et al. (2023)	✓							0.49	0.29	-
Søren Kierkegaard	Talavera Cepeda et al. (2023)	✓		✓					0.49	-	-
Prodicus	Moosavi M. and Eetemadi (2023)	✓			✓	🐙			0.48	0.30	-
MaChAmp (Robert S. Hartman)	van der Goot (2023)	✓				🐙			0.48	0.34	0.19
Andronicus of Rhodes	Papadopoulos et al. (2023)	✓				🐙			0.48	-	-
Rudolf Christoph Eucken	Saha and Srihari (2023)	✓	✓	✓		🐙			0.48	-	-
Aristoxenus	Zaikis et al. (2023)	✓	✓			🐙			0.47	0.25	-
Noam Chomsky	Honda and Wilharm (2023)	✓			✓	🐙			0.47	0.26	-
Tom Regan	Koichi Tanigaki	✓							0.47	-	-
Sina (Seyyed Hossein Nasr)	Ghahroodi et al. (2023)	✓				🐙	🐳		0.47	0.25	0.24
LRL_NC (George Boole)	Tandon and Chatterjee (2023)	✓				🐙			0.46	0.27	-
Epicurus	Fang et al. (2023)	✓	✓		✓	🐙			0.46	0.27	-
I2C Huelva (Marquis de Sade)	El Balima Cordero et al. (2023)	✓							0.46	-	-
Lauri Ingman	Paulissen and Wendt (2023)	✓				🐙			0.44	-	-
Augustine of Hippo	Ferrara et al. (2023)	✓	✓				🐳		0.44	0.23	0.19
Mary Daly	Dong Qing	✓							0.43	-	-
Niccolò Machiavelli	Anonymous	✓					🐳		0.42	0.23	0.25
Philippa Foot	Anonymous	-	-	-	-				0.42	-	-
BERT (Aristotle)	Kiesel et al. (2022)	✓				🐙	🐳		0.42	0.28	0.24
Francis Bacon	Hasanaliyev et al. (2023)	✓	✓			🐙			0.42	-	-
Tenzin Gyatso	Kandru et al. (2023)	✓				🐙			0.41	-	-
TeamEC (Johann Georg Walch)	Stefanovitch et al. (2023)								0.40	-	-
StFX-NLP (Jesus of Nazareth)	Heavey et al. (2023)	✓	✓	✓		🐙			0.40	-	-
Quintilian	Mopidevi and Chenna (2023)	✓		✓					0.38	0.20	0.10
Joseph Fletcher	Ewelina Gajewska								0.34	-	-
Francisco de Vitoria	Umberto Altieri	✓				🐙			0.32	-	-
1-Baseline (Aristotle)	Kiesel et al. (2022)					🐙	🐳		0.26	0.13	0.15
Martha Nussbaum	Anonymous	✓							0.24	-	-
Johann Friedrich Herbart	Anonymous	✓					🐳		0.18	-	-
Friedrich Nietzsche	Sundharram et al. (2023)		✓			🐙			0.01	-	-

Table 3: Overview of the teams who participated in ValueEval'23, along with our two baselines (BERT and 1-Baseline) and a meta-approach that takes the best approach for each single value category (in gray). If the team handed in a paper to SemEval'23, the paper is cited. Shows for every team whether they used at least one approach that is based on transformers (Tr.), an ensemble (En.), formulated the task not as a classification problem (NC.), or employed value descriptions/semantics (Val.). Teams Confucius and Philippa Foot provided no information on their approach. If available, the table also contains hyperlinks to the team's source code (🐙), Docker image (🐳), and web demo (🌐). Also shows the approaches' scores on the test datasets: Main, Nahj al-Balagha (Nahj.), and New York Times (NYT).

sisting of premise, conclusion, and stance, were combined in several ways in order to create four different input datasets. For example, in one dataset, sample labels were merged in order to equalize data imbalance. Each trained transformer has classifiers for all labels. The final results were ensembled by weight voting based on the F_1 score in the validation set of each value category. Several loss functions were considered, where a class-balanced loss combined with a negative tolerant regularization proved to be the best approach. Also, different classification thresholds were tried but without performance improvement.

Employed Models Transformer-based models were the most dominant techniques (35/37; cf. Table 3). All submissions that used transformers fine-tuned them on data except team StFX-NLP (Jesus of Nazareth), who integrated BERT's embeddings as features into their approach. Only two teams relied on classical machine learning models such as decision trees and SVMs (team Friedrich Nietzsche) or logistic regression (team Joseph Fletcher). Among the 35 transformer-based submissions, four approaches considered reformulating the task as a non-classification task, while 31 approaches performed direct classification on the provided labels.

Non-classification Approaches Four teams formulated the task as natural language inference: Søren Kierkegaard, Rudolf Christoph Eucken, Jesus of Nazareth, and Quintilian). For example, Team Søren Kierkegaard constructed samples from the premise and the corresponding value category with entailment or contradictory labels inferred from whether the premise has the value category. Then, they fine-tuned a RoBERTa model that was already pre-trained on the natural language inference task (Bowman et al., 2015). One team formulated the task as question-answering: Team Hitachi (R. M. Hare) used BART and T5 as backbone models. The predictions were made by feeding the model a yes/no question for each value category given a (premise, stance, conclusion) triple. They also experimented with formulating the task as zero-shot question-answering with chain-of-thought prompting (Kojima et al., 2022) using GPT-3 (text-davinci-003).

Employed Data Out of the 37 submissions, all used the premise as an input, 25 used the conclusion, and 24 used the stance. For example, team Tübingen (Stanley Grenz) found that utilizing the stance as part of the output to predict instead of being in the input led to better effectiveness. Four teams augmented the data with our definitions of the value categories (4 approaches): team Prodicus created extra instances by concatenating human values and their descriptions as input and the corresponding value category as a label; team Noam Chomsky used the similarity between premises and the corresponding human values as extra features. Team Hitachi also employed value category descriptions and formulated the task as a question-answering problem, as described above. Team Epicurus augments their data by adding definitional statements formulated based on several human value surveys, such as the world value survey (Haerpfer et al., 2020), and Rokeach value survey (Rokeach, 1973).

Special Training Besides team Adam Smith, as described above, three more teams pre-trained on other corpora: team Niccolo Machiavelli fine-tuned DeBERTaV3, and team Hitachi fine-tuned RoBERTa on ValueNet (Qiu et al., 2022), while team MaChAmp (Robert S. Hartman) fine-tuned on data from the other SemEval'23 tasks ("intermediate training"). However, for most SemEval tasks—including this one—the intermediate training did not result in higher effectiveness. Two teams employed contrastive learning: Mao Zedong and T. M. Scanlon). Team SUTNLP (David Gauthier) employed adversarial training to improve the generalization of their model.

Label Semantics Five submissions (Mao Zedong, SUTNLP, Hitachi, Prodicus, Epicurus) employed label semantics by learning it from data or using the definitions of human values. For example, team Mao-Zedong utilized a label-guided attention mechanism to learn label-specific representations from the input. They also utilize a contrastive loss function to pull instances with similar labels together. Team SUTNLP learns a semantic representation of labels by encoding them jointly with their corresponding premises and then employs adversarial training to enhance the generalization of their model. They further capture label correlations by constructing a graph over labels using a Graph Convolutional Network (Zhang et al., 2019), where nodes (hidden label representations) are connected if there is at least one argument in the training data that contains these labels.

6 Results

We evaluated the submitted approaches on the Touchè23-ValueEval dataset, both the main and supplementary parts (Section 6.1). Though we allowed teams to submit predictions only for a subset of the 20 value categories, only one team did so, and not on their best run. We use two baselines from our previous work (Kiesel et al., 2022):

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

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

Additionally, we report on a meta-classification study for an analysis of straightforward approach-combinations (Section 6.2) and describe our efforts to disseminate the results (Section 6.3).

6.1 Results of Participants

Table 3 shows the best macro F_1-score each team achieved for each of the three test datasets. Team Adam Smith reached the highest F_1-score for the main and Nahj al-Balagha test sets, whereas team Hitachi (R. M. Hare) was best on the New York Times dataset. The table shows a considerable increase over the BERT baseline for all datasets: 0.56 vs. 0.42 for main, 0.40 vs. 0.28 for Nahj al-Balagha, and 0.34 vs. 0.24 for New York Times. The naive 1-baseline performs even worse (0.26, 0.13, and 0.15), though the F_1-score favors it in comparison with accuracy or other measures. Though the scores for the supplementary datasets are lower, one has to take into consideration that the teams had no training data from these datasets. In light of the very different nature of these datasets compared to the main dataset, we find these results to be very encouraging for building robust systems.

However, not every approach performed equally well for all value categories. Table 3 shows the "best per value category" system at the top, which is a hypothetical system that uses for each value category the run that performed best for that category—among all submitted runs, not just the ones that achieved a best macro F_1-score. The improvement of this hypothetical system over the best single runs is small but noticeable for the main test set (0.59 vs. 0.56), but quite large for Nahj al-Balagha (0.48 vs. 0.40) and even larger for New York Times (0.47 vs. 0.34). For more details, Tables 5 and 6 in the appendix show the F_1-score of the best approaches per team for each value category. For example, the best approach for *stimulation* and *hedonism* outperforms the approach by Adam Smith considerably (0.39 for both vs. 0.22 and 0.29), but does not perform well overall (it is not in the list of best approaches per team).

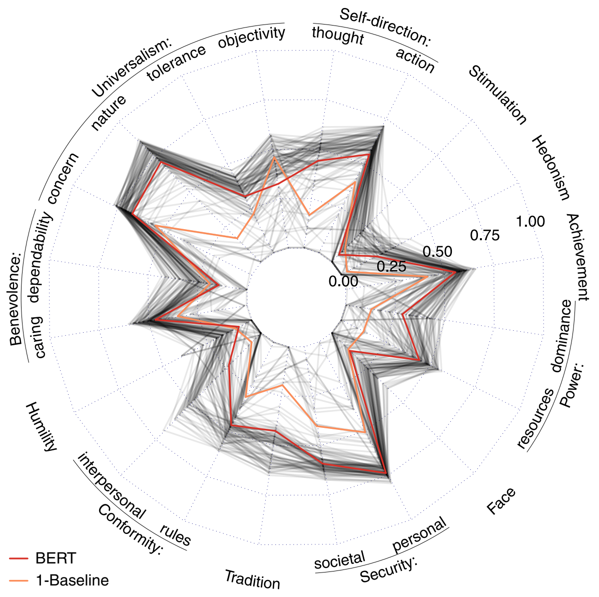


Figure 4: F_1-score for each value category of each submission to the main test dataset, with lines corresponding to one submission each. Baselines are colored.

For a visual impression of the submitted runs' performance, the radar plot in Figure 4 shows the F_1-score of every submitted run for each value category, where each run corresponds to one line. The plot shows that most runs did improve over the baselines for most value categories: there is a "dark band" that is consistently outside of the red "BERT" line. Moreover, the plot reveals some peculiarities, like that only few runs are at the top for *stimulation*, *hedonism*, *face*, *conformity: interpersonal* and *humility*. Since *stimulation* and *hedonism* are next to each other, one can also see that the top run in each one of these performed worse on the respective other category. The plot also reveals the anomaly of the *universalism: objectivity* category, for which the 1-baseline performed considerably better that the BERT baseline and many submitted runs. This is the only category that was added—based on a comparison with other value taxonomies (Kiesel et al., 2022)—to the 19 categories of the Schwartz taxonomy. Schwartz et al. (2012) refrained from including such a category as they found truth-seeking to be correlated with several other categories on all sides of the wheel. This observation might also be connected to the anomaly we observe in the figure.

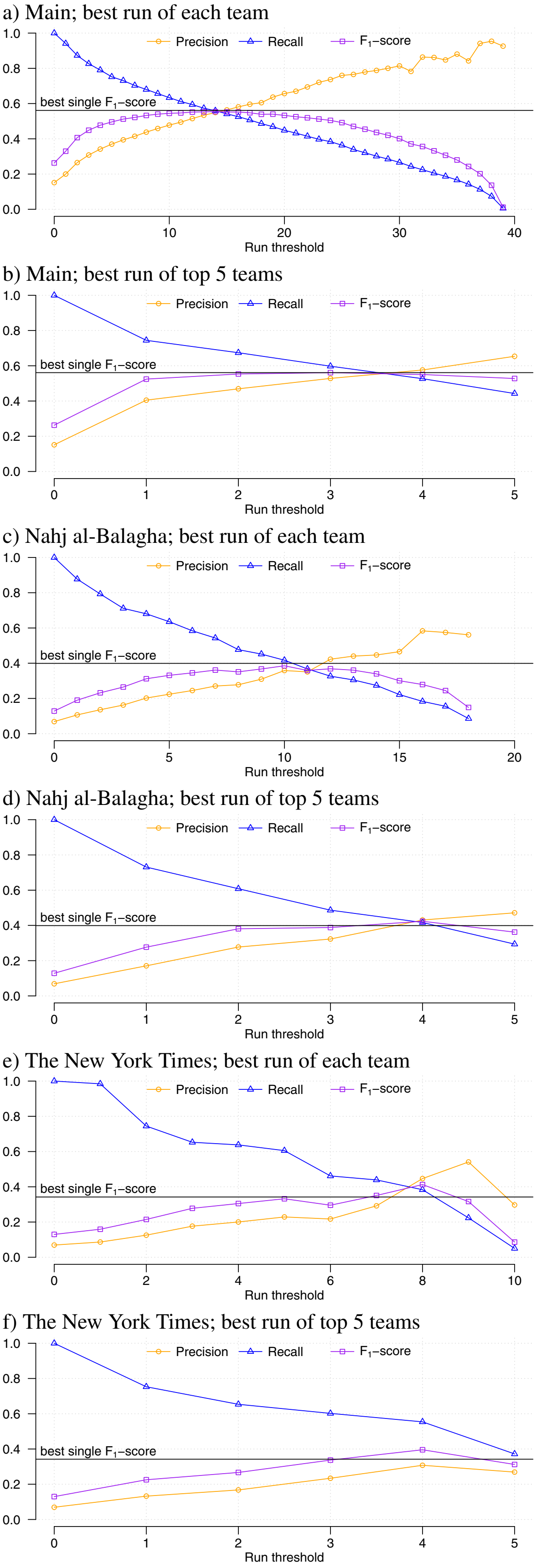
6.2 Results from Meta-classification

As a form of meta-analysis we tested the use of basic ensemble methods on top of the submitted run files (excluding baselines). Figure 5 shows the effectiveness of a simple voting scheme: the ensemble classifies an argument as resorting to a value category if the same was done in at least X of the submitted runs ("run threshold"). Thus, if the threshold is 0, the voting scheme corresponds to the 1-Baseline. We employ either the best run for each team (per dataset, as determined by macro F_1-score), or only the top-5 runs of these. We refrained from using more involved ensemble techniques as this analysis is of theoretical nature either way: the computational costs of such an ensemble makes its use unfeasible in practice.

Unlike in a previous shared task of ours (Kiesel et al., 2019), the voting scheme does not (main test dataset, Figure 5 a,b) or only marginally (supplementary test datasets, Figure 5 c–f) improve over the best single run (Table 3, solid line in Figure 5).

The achieved F_1-score is similar to the best single run for relatively large ranges of the run threshold, for which every increase in precision is traded with a similar decrease in recall. For top-5 runs on the main test set (Figure 5 b) this equilibrium even holds for all run thresholds except 0, which corresponds to a even field among the top 5 teams. The decrease in recall from 74% to 44% shows that, though many argument-value category pairs seem to be clear in that they were found unanimous among the top teams (44%), the approaches are diverse enough to detect many more pairs. This situation is more extreme for the Nahj al-Balagha set (73% to 29%; Figure 5 d) but similar for the New York Times set (75% to 37%; Figure 5 f).

The extreme values for the run threshold show both the very difficult cases—ratio of argument-value category pairs which were found in none of the runs (1 - recall for run threshold of 1)—and very easy cases—ratio of pairs found in each run (recall at highest run threshold). The ratio of difficult (easy) cases is 6% (<1%), 12% (9%), and 2% (5%) for main, Nahj al-Balagha, and New York Times, respectively (Figure 5 a,c,e). Though these numbers are not fully comparable given the varying number of submitting teams, they highlight that most pairs have indeed been found by at least one of the submitted (best) runs.


a) Main; best run of each team

b) Main; best run of top 5 teams

c) Nahj al-Balagha; best run of each team

d) Nahj al-Balagha; best run of top 5 teams

e) The New York Times; best run of each team

f) The New York Times; best run of top 5 teams

Figure 5: Evaluation of a meta-classifier that detects a value category if at least a number of runs above some threshold do so, on each test dataset and using either the best run of each team or of the top-5 teams only.

6.3 Results Dissemination

Table 3 cites all submitted papers and links to the source code and Docker images of the teams. Moreover, we worked together with the top-ranked team, Adam Smith, to provide an online demo and executable of their best-performing approach.[12]

7 Conclusion

This paper reports on the setup, participation, results, and insights gained from the Touché task on identifying human values behind arguments (ValueEval'23), hosted as Task 4 at SemEval-2023. We detail the construction both of the main dataset of 8865 arguments, as well as the supplementary dataset of 459 arguments used to test model generalizability. Moreover, we provide a systematic overview of the 29 papers submitted by the participants, compare their approaches, and perform an ad-hoc meta classification.

Through the use of TIRA (Fröbe et al., 2023) we can directly evaluate the approaches of the 8 teams who used Docker submission on new datasets for human values detection, provided they are formatted like the datasets presented here. Moreover 23 teams set up an open source repository for making their code available to other researchers.

Given the implicit nature of human values, very promising results were achieved during the task, with F_1-scores above 0.56 on the main test set and up to 0.87 for individual categories. However, many challenges remain, as evidenced by the low peak F_1-score of 0.39 for stimulation, hedonism, face, and humility. The results show that, like in many other NLP tasks, transformer models could be used to great effect, and in particular ensembling approaches showed to be quite effective. The results between the main and supplementary datasets were quite different as anticipated due to their dissimilarity, yet teams managed to beat the baselines on the supplementary datasets by a wide margin despite the total lack of in-domain training data. Moreover, several teams reported that using our supplementary validation set as an addition to the training set boosted results, suggesting that a larger dataset of this kind will probably assist in improving the performance of future models even further. Thus, there is still room for improvement on this task, and whether progress will come from novel models or creative applications of existing techniques remains to be seen.

[12] https://values.args.me/

Acknowledgements

Our thanks go out to all participating teams; your contributions made this task a success. Our special thanks go out to the SemEval organizers and to the TIRA team for their constant support.

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-IJCNLP'19)*, pages 2922–2932. ACL.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

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.

Abdul Aziz, MD. Akram Hossain, and Abu Nowshed Chy. 2023. CSECU-DSG at SemEval-2023 Task 4: Fine-tuning DeBERTa Transformer Model with Cross-fold Training and Multi-sample Dropout for Human Values Identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2022–2028, Toronto, Canada. Association for Computational Linguistics.

Georgios Balikas. 2023. John-Arthur at SemEval-2023 Task 4: Fine-Tuning Large Language Models for Arguments Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1461–1465, Toronto, Canada. Association for Computational Linguistics.

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 (AACL'22)*, pages 418–422, 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)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Fidan Can. 2023. Tübingen at SemEval-2023 Task 4: What Can Stance Tell? A Computational Study on Detecting Human Values behind Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1796–1801, Toronto, Canada. 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.

Nordin El Balima Cordero, Jacinto Mata, Victoria Pachón, and Abel Pichardo Estevez. 2023. I2C Huelva at SemEval-2023 Task 4: A Resampling and Transformers Approach to Identify Human Values behind Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1415–1420, Toronto, Canada. Association for Computational Linguistics.

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

Christian Fang, Qixiang Fang, and Dong Nguyen. 2023. Epicurus at SemEval-2023 Task 4: Improving Prediction of Human Values behind Arguments by Leveraging Their Definitions. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 223–231, Toronto, Canada. Association for Computational Linguistics.

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.

Alfio Ferrara, Sergio Picascia, and Elisabetta Rocchetti. 2023. Augustine of Hippo at SemEval-2023 Task 4: An Explainable Knowledge Extraction Method to Identify Human Values in Arguments with SuperASKE. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1073–1082, Toronto, Canada. Association for Computational Linguistics.

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.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR'23)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Omid Ghahroodi, Mohammad Ali Sadraei Javaheri, Doratossadat Dastgheib, Mahdieh Soleymani Baghshah, Mohammad Hossein Rohban, Hamid R. Rabiee, and Ehsaneddin Asgari. 2023. Sina at SemEval-2023 Task 4: A Class-Token Attention-based Model for Human Value Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2198–2201, Toronto, Canada. 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 Cnalysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, pages 7805–7813.

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.

Kenan Hasanaliyev, Kevin Li, Saanvi Chawla, Michael Nath, Rohan Sanda, Justin Wu, William Huang, Daniel Yang, Shane Mion, and Kiran Bhat. 2023. Francis Bacon at SemEval-2023 Task 4: Ensembling BERT and GloVe for Value Identification in Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2073–2076, Toronto, Canada. Association for Computational Linguistics.

Ethan Heavey, Milton King, and James Hughes. 2023. StFX-NLP at SemEval-2023 Task 4: Unsupervised and Supervised Approaches to Detecting Human Values in Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 205–211, Toronto, Canada. Association for Computational Linguistics.

Hamed Hematian Hemati, Sayed Hesam Alavian, Hossein Sameti, and Hamid Beigy. 2023. SUTNLP at SemEval-2023 Task 4: LG-Transformer for Human Value Detection. In *Proceedings of the 17th*

International Workshop on Semantic Evaluation (SemEval'23), pages 342–348, Toronto, Canada. Association for Computational Linguistics.

Sumire Honda and Sebastian Wilharm. 2023. Noam Chomsky at SemEval-2023 Task 4: Hierarchical Similarity-aware Model for Human Value Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1392–1397, Toronto, Canada. Association for Computational Linguistics.

Siri Venkata Pavan Kumar Kandru, Bhavyajeet Singh, Ankita Maity, Kancharla Aditya Hari, and Vasudeva Varma. 2023. Tenzin-Gyatso at SemEval-2023 Task 4: Identifying Human Values behind Arguments Using DeBERTa. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2096–2100, Toronto, Canada. Association for Computational Linguistics.

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'22)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19)*, pages 829–839. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems (NeurIPS'22)*.

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.

Long Ma. 2023. PAI at SemEval-2023 Task 4: A General Multi-label Classification System with Class-balanced Loss Function and Ensemble Module. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 258–263, Toronto, Canada. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Milad Molazadeh Oskuee, Mostafa Rahgouy, Hamed Babaei Giglou, and Cheryl D. Seals. 2023. T.M. Scanlon at SemEval-2023 Task 4: Leveraging Pretrained Language Models for Human Value Argument Mining with Contrastive Learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 611–616, Toronto, Canada. Association for Computational Linguistics.

Erfan Moosavi M. and Sauleh Eetemadi. 2023. Prodicus at SemEval-2023 Task 4: Enhancing Human Value Detection with Data Augmentation and Fine-Tuned Language Models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2067–2072, Toronto, Canada. Association for Computational Linguistics.

Ajay Narasimha Mopidevi and Hemanth Chenna. 2023. Quintilian at SemEval-2023 Task 4: Grouped BERT for Multi-Label Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1642–1645, Toronto, Canada. Association for Computational Linguistics.

Georgios Papadopoulos, Marko Kokol, Maria Dagioglou, and Georgios Petasis. 2023. Andronicus of Rhodes at SemEval-2023 Task 4: Transformer-Based Human Value Detection Using Four Different Neural Network Architectures. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 550–556, Toronto, Canada. Association for Computational Linguistics.

Spencer Paulissen and Caroline Wendt. 2023. Lauri Ingman at SemEval-2023 Task 4: A Chain Classifier for Identifying Human Values behind Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 193–198, Toronto, Canada. Association for Computational Linguistics.

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.

Sougata Saha and Rohini Srihari. 2023. Rudolf Christoph Eucken at SemEval-2023 Task 4: An Ensemble Approach for Identifying Human Values from Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 674–677, Toronto, Canada. Association for Computational Linguistics.

Mario Scharfbillig, Vladimir Ponizovskiy, Zsuzsanna Pasztor, Julian Keimer, and Giuseppe Tirone. 2022. Monitoring Social Values in Online Media Articles on Child Vaccinations. Technical Report KJ-NA-31-324-EN-N, European Commission's Joint Research Centre, Luxembourg.

Mario Scharfbillig, Laura Smillie, David Mair, Marta Sienkiewicz, Julian Keimer, Raquel Pinho Dos Santos, Hélder Vinagreiro Alves, Elise Vecchione, and Laurenz Scheunemann. 2021. Values and Identities - a Policymaker's Guide. Technical Report KJ-NA-30800-EN-N, European Commission's Joint Research Centre, Luxembourg.

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'23)*, pages 540–549, 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).

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

Xianxian Song, Jinhui Zhao, Ruiqi Cao, Linchi Sui, Binyang Li, and Tingyue Guan. 2023. Arthur Caplan at SemEval-2023 Task 4: Enhancing Human Value Detection through Fine-tuned Pre-trained Models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1987–1993, Toronto, Canada. Association for Computational Linguistics.

Nicolas Stefanovitch, Bertrand de Longueville, and Mario Scharfbillig. 2023. TeamEC at SemEval-2023 Task 4: Transformers vs. Low-Resource Dictionaries, Expert Dictionary vs. Learned Dictionary. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2141–2145, Toronto, Canada. Association for Computational Linguistics.

Sruthi Sundharram, Abdul Jawad Mohammed, and Sanidhya Sharma. 2023. Friedrich Nietzsche at SemEval-2023 Task 4: Detection of Human Values from Text Using Machine Learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2213–2217, Toronto, Canada. Association for Computational Linguistics.

Ignacio Talavera Cepeda, Amalie Brogaard Pauli, and Ira Assent. 2023. Søren Kierkegaard at SemEval-2023 Task 4: Label-aware Text Classification Using Natural Language Inference. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1905–1911, Toronto, Canada. Association for Computational Linguistics.

Kushagri Tandon and Niladri Chatterjee. 2023. LRL_NC at SemEval-2023 Task 4: The Touche23-George-boole Approach for Multi-Label Classification of Human-Values behind Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 136–142, Toronto, Canada. Association for Computational Linguistics.

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.

Masaya Tsunokake, Atsuki Yamaguchi, Yuta Koreeda, Hiroaki Ozaki, and Yasuhiro Sogawa. 2023. Hitachi at SemEval-2023 Task 4: Exploring Various Task Formulations Reveals the Importance of Description Texts on Human Values. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1756–1768, Toronto, Canada. Association for Computational Linguistics.

Rob van der Goot. 2023. MaChAmp at SemEval-2023 Tasks 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12: On the Effectiveness of Intermediate Training on an Uncurated Collection of Datasets. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 232–247, Toronto, Canada. 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'09)*, volume 6057 of *Lecture Notes in Computer Science*, pages 79–93. Springer.

Dimitrios Zaikis, Stefanos D. Stefanidis, Konstantinos Anagnostopoulos, and Ioannis Vlahavas. 2023. Aristoxenus at SemEval-2023 Task 4: A Domain-Adapted Ensemble Approach to the Identification of Human Values behind Arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1066–1072, Toronto, Canada. Association for Computational Linguistics.

Che Zhang, Ping'an Liu, Zhenyang Xiao, and Haojun Fei. 2023. Mao-Zedong at SemEval-2023 Task 4: Label Represention Multi-Head Attention Model with Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 435–441, Toronto, Canada. Association for Computational Linguistics.

Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph Convolutional Networks: A Comprehensive Review. *Computational Social Networks*, 6(1):1–23.

A Appendix

Value category descriptions (Table 4) and results per value (Tables 5 and 6).

Value category (level 2)	**Contained values (level 1)**
○ *Self-direction: thought* It is good to have own ideas and interests.	*Be creative*: arguments towards more creativity or imagination *Be curious*: arguments towards more curiosity, discoveries, or general interestingness *Have freedom of thought*: arguments toward people figuring things out on their own, towards less censorship, or towards less influence on thoughts
○ *Self-direction: action* It is good to determine one's own actions.	*Be choosing own goals*: arguments towards allowing people to choose what is best for them, to decide on their life, and to follow their dreams *Be independent*: arguments towards allowing people to plan on their own and not ask for consent *Have freedom of action*: arguments towards allowing people to be self-determined and able to do what they want *Have privacy*: arguments towards allowing for private spaces, time alone, and less surveillance, or towards more control on what to disclose and to whom
○ *Stimulation* It is good to experience excitement, novelty, and change.	*Have an exciting life*: arguments towards allowing people to experience foreign places and special activities or having perspective-changing experiences *Have a varied life*: arguments towards allowing people to engage in many activities and change parts of their life or towards promoting local clubs (sports, ...) *Be daring*: arguments towards more risk-taking
○ *Hedonism* It is good to experience pleasure and sensual gratification.	*Have pleasure*: arguments towards making life enjoyable or providing leisure, opportunities to have fun, and sensual gratification
○ *Achievement* It is good to be successful in accordance with social norms.	*Be ambitious*: arguments towards allowing for ambitions and climbing up the social ladder *Have success*: arguments towards allowing for success and recognizing achievements *Be capable*: arguments towards acquiring competence in certain tasks, being more effective, and showing competence in solving tasks *Be intellectual*: arguments towards acquiring high cognitive skills, being more reflective, and showing intelligence *Be courageous*: arguments towards being more courageous and having people stand up for their beliefs
○ *Power: dominance* It is good to be in positions of control over others.	*Have influence*: arguments towards having more people to ask for a favor, more influence, and more ways to control events *Have the right to command*: arguments towards allowing the right people to take command, putting experts in charge, and clearer hierarchies of command, or towards fostering leadership
○ *Power: resources* It is good to have material possessions and social resources.	*Have wealth*: arguments towards allowing people to gain wealth and material possession, show their wealth, and exercise control through wealth, or towards financial prosperity
○ *Face* It is good to maintain one's public image.	*Have social recognition*: arguments towards allowing people to gain respect and social recognition or avoid humiliation *Have a good reputation*: arguments towards allowing people to build up their reputation, protect their public image, and spread reputation
○ *Security: personal* It is good to have a secure immediate environment.	*Have a sense of belonging*: arguments towards allowing people to establish, join, and stay in groups, show their group membership, and show that they care for each other, or towards fostering a sense of belonging *Have good health*: arguments towards avoiding diseases, preserving health, or having physiological and mental well-being *Have no debts*: arguments towards avoiding indebtedness and having people return favors *Be neat and tidy*: arguments towards being more clean, neat, or orderly *Have a comfortable life*: arguments towards providing subsistence income, having no financial worries, and having a prosperous life, or towards resulting in a higher general happiness
○ *Security: societal* It is good to have a secure and stable wider society.	*Have a safe country*: arguments towards a state that can better act on crimes, and defend or care for its citizens, or towards a stronger state in general *Have a stable society*: arguments towards accepting or maintaining the existing social structure or towards preventing chaos and disorder at a societal level
○ *Tradition* It is good to maintain cultural, family, or religious traditions.	*Be respecting traditions*: arguments towards allowing to follow one's family's customs, honoring traditional practices, maintaining traditional values and ways of thinking, or promoting the preservation of customs *Be holding religious faith*: arguments towards allowing the customs of a religion and to devote one's life to their faith, or towards promoting piety and the spreading of one's religion

Table 4 (continued on next page).

Table 4 (continued).

Value category (level 2)	Contained values (level 1)
○ *Conformity: rules* It is good to comply with rules, laws, and formal obligations.	*Be compliant*: arguments towards abiding to laws or rules and promoting to meet one's obligations or recognizing people who do *Be self-disciplined*: arguments towards fostering to exercise restraint, follow rules even when no-one is watching, and to set rules for oneself *Be behaving properly*: arguments towards avoiding to violate informal rules or social conventions or towards fostering good manners
○ *Conformity: interpersonal* It is good to avoid upsetting or harming others.	*Be polite*: arguments towards avoiding to upset other people, taking others into account, and being less annoying for others *Be honoring elders*: arguments towards following one's parents or showing faith and respect towards one's elders
○ *Humility* It is good to recognize one's own insignificance in the larger scheme of things.	*Be humble*: arguments towards demoting arrogance, bragging, and thinking one deserves more than other people, or towards emphasizing the successful group over single persons and giving back to society *Have life accepted as is*: arguments towards accepting one's fate, submitting to life's circumstances, and being satisfied with what one has
○ *Benevolence: caring* It is good to work for the welfare of one's group's members.	*Be helpful*: arguments towards helping the people in one's group and promoting to work for the welfare of others in one group *Be honest*: arguments towards being more honest and recognizing people for their honesty *Be forgiving*: arguments towards allowing people to forgive each other, giving people a second chance, and being merciful, or towards providing paths to redemption *Have the own family secured*: arguments towards allowing people to have, protect, and care for their family *Be loving*: arguments towards fostering close relationships and placing the well-being of others above the own, or towards allowing to show affection, compassion, and sympathy
○ *Benevolence: dependability* It is good to be a reliable and trustworthy member of one's group.	*Be responsible*: arguments towards clear responsibilities, fostering confidence, and promoting reliability *Have loyalty towards friends*: arguments towards being a dependable, trustworthy, and loyal friend, or towards allowing to give friends a full backing
○ *Universalism: concern* It is good to strive for equality, justice, and protection for all people.	*Have equality*: arguments towards fostering people of a lower social status, helping poorer regions of the world, providing all people with equal opportunities in life, and resulting in a world were success is less determined by birth *Be just*: arguments towards allowing justice to be 'blind' to irrelevant aspects of a case, promoting fairness in competitions, protecting the weak and vulnerable in society, and resulting a world were people are less discriminated based on race, gender, and so on, or towards fostering a general sense for justice *Have a world at peace*: arguments towards nations ceasing fire, avoiding conflicts, and ending wars, or promoting to see peace as fragile and precious or to care for all of humanity
○ *Universalism: nature* It is good to preserve the natural environment.	*Be protecting the environment*: arguments towards avoiding pollution, fostering to care for nature, or promoting programs to restore nature *Have harmony with nature*: arguments towards avoiding chemicals and genetically modified organisms (especially in nutrition), or towards treating animals and plants like having souls, promoting a life in harmony with nature, and resulting in more people reflecting the consequences of their actions towards the environment *Have a world of beauty*: arguments towards allowing people to experience art and stand in awe of nature, or towards promoting the beauty of nature and the fine arts
○ *Universalism: tolerance* It is good to accept and try to understand those who are different from oneself.	*Be broadminded*: arguments towards allowing for discussion between groups, clearing up with prejudices, listening to people who are different from oneself, and promoting to life within a different group for some time, or towards promoting tolerance between all kinds of people and groups in general *Have the wisdom to accept others*: arguments towards allowing people to accept disagreements and people even when one disagrees with them, to promote a mature understanding of different opinions, or to decrease partisanship or fanaticism
○ *Universalism: objectivity* It is good to search for the truth and think in a rational and unbiased way.	*Be logical*: arguments towards going for the numbers instead of gut feeling, towards a rational, focused, and consistent way of thinking, towards a rational analysis of circumstances, or towards promoting the scientific method *Have an objective view*: arguments towards fostering to seek the truth, to take on a neutral perspective, to form an unbiased opinion, and to weigh all pros and cons, or towards providing people with the means to make informed decisions

Table 4: Descriptions for each of the 20 employed value categories (level 2) and their respective contained values (level 1; 54 total). These value descriptions were also used (in slightly adapted form as bullet points) in the dataset annotation and are distributed along with the dataset.

Team	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Best per value category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Adam Smith	.56	.59	.71	.22	.29	.66	.48	.52	.30	.79	.67	.65	.61	.61	.19	.60	.36	.74	.84	.41	.53
John Arthur	.55	.56	.70	.27	.25	.65	.50	.52	.39	.76	.60	.63	.60	.69	.24	.55	.41	.74	.86	.44	.58
PAI (Theodor Zwinger)	.54	.59	.71	.29	.32	.61	.45	.49	.36	.79	.67	.55	.59	.58	.12	.58	.34	.76	.85	.42	.48
Mao Zedong	.53	.53	.70	.26	.29	.60	.45	.54	.31	.77	.65	.58	.60	.51	.16	.59	.42	.73	.85	.43	.55
Confucius	.53	.52	.71	.25	.32	.61	.44	.53	.39	.75	.64	.61	.57	.47	.16	.58	.40	.73	.84	.42	.57
Arthur Caplan	.51	.53	.65	.26	.30	.62	.43	.52	.29	.73	.62	.61	.56	.48	.16	.54	.34	.72	.80	.40	.54
Hitachi (R. M. Hare)	.51	.48	.66	.22	.23	.61	.43	.45	.32	.74	.63	.57	.54	.47	.15	.53	.36	.74	.81	.42	.55
SUTNLP (David Gauthier)	.50	.55	.67	.18	.18	.60	.30	.53	.31	.74	.60	.56	.46	.55	.32	.52	.30	.76	.72	.43	.45
Tübingen (Stanley Grenz)	.50	.55	.67	.10	.29	.61	.34	.49	.18	.77	.65	.62	.52	.29	.12	.57	.23	.75	.79	.38	.42
Georg Simmel	.50	.53	.64	.21	.29	.59	.36	.48	.26	.76	.63	.47	.57	.57	.22	.54	.36	.72	.79	.44	.43
T. M. Scanlon	.49	.56	.67	.18	.39	.63	.36	.48	.26	.75	.63	.47	.53	.38	.20	.50	.31	.73	.82	.37	.42
CSECU-DSG (Fazlur Rahman)	.49	.54	.69	.12	.26	.60	.32	.48	.02	.77	.66	.64	.53	.29	.08	.55	.28	.78	.82	.37	.51
Søren Kierkegaard	.49	.53	.58	.19	.30	.58	.35	.50	.27	.75	.62	.59	.53	.58	.18	.54	.15	.73	.77	.38	.39
Prodicus	.48	.53	.61	.07	.27	.54	.32	.41	.15	.73	.62	.54	.51	.35	.11	.53	.15	.73	.78	.37	.43
MaChAmp (Robert S. Hartman)	.48	.52	.62	.22	.23	.59	.41	.45	.30	.75	.62	.48	.52	.49	.20	.53	.24	.72	.81	.36	.38
Andronicus of Rhodes	.48	.47	.65	.25	.29	.58	.35	.54	.30	.71	.60	.51	.54	.27	.14	.52	.31	.69	.76	.39	.48
Rudolf Christoph Eucken	.48	.40	.60	.20	.23	.60	.41	.48	.25	.70	.57	.61	.51	.46	.13	.49	.24	.71	.80	.26	.42
Aristoxenus	.47	.58	.66	.09	.25	.58	.07	.50	.29	.75	.61	.56	.51	.52	.27	.49	.20	.76	.77	.34	.40
Noam Chomsky	.47	.51	.59	.15	.28	.59	.36	.47	.22	.72	.61	.48	.56	.36	.15	.51	.23	.71	.78	.40	.41
Tom Regan	.47	.44	.63	.17	.25	.58	.18	.47	.22	.72	.62	.52	.47	.47	.29	.50	.28	.74	.66	.27	.36
Sina (Seyyed Hossein Nasr)	.47	.42	.61	.20	.21	.61	.39	.55	.24	.73	.58	.45	.54	.52	.18	.50	.24	.72	.77	.38	.49
LRL_NC (George Boole)	.46	.49	.61	.05	.20	.61	.28	.47	.23	.74	.61	.49	.49	.27	.19	.53	.14	.71	.77	.34	.41
Epicurus	.46	.49	.59	.22	.33	.57	.36	.50	.23	.70	.61	.47	.45	.26	.19	.47	.28	.68	.74	.34	.52
I2C Huelva (Marquis de Sade)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
Lauri Ingman	.44	.50	.62	.05	.13	.54	.16	.36	.21	.71	.54	.49	.47	.32	.33	.46	.28	.69	.68	.31	.37
Augustine of Hippo	.44	.40	.58	.22	.09	.58	.33	.51	.20	.71	.59	.45	.50	.32	.20	.47	.28	.69	.72	.33	.45
Mary Daly	.43	.41	.57	.14	.17	.57	.29	.51	.23	.65	.53	.48	.48	.31	.17	.42	.29	.64	.72	.36	.51
Niccolò Machiavelli	.42	.52	.63	.05	.14	.56	.32	.37	.00	.70	.61	.48	.43	.00	.11	.43	.20	.74	.70	.24	.37
Philippa Foot	.42	.46	.60	.05	.26	.50	.17	.42	.13	.72	.59	.53	.42	.30	.14	.48	.10	.73	.60	.24	.35
BERT (Aristotle)	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
Francis Bacon	.42	.40	.53	.19	.28	.57	.28	.49	.17	.67	.54	.48	.47	.24	.11	.41	.27	.64	.68	.34	.50
Tenzin Gyatso	.41	.44	.62	.07	.12	.47	.33	.45	.10	.70	.57	.45	.42	.21	.12	.47	.21	.67	.73	.31	.37
TeamEC (Johann Georg Walch)	.40	.42	.50	.04	.13	.53	.17	.36	.10	.68	.53	.47	.48	.21	.17	.46	.25	.68	.56	.34	.37
StFX-NLP (Jesus of Nazareth)	.40	.40	.52	.07	.17	.47	.30	.40	.05	.71	.55	.41	.48	.07	.07	.44	.17	.63	.69	.36	.39
Quintilian	.38	.49	.58	.00	.00	.58	.23	.44	.00	.66	.52	.47	.49	.00	.00	.41	.30	.65	.64	.38	.45
Joseph Fletcher	.34	.43	.45	.00	.00	.40	.05	.28	.09	.69	.53	.46	.42	.44	.05	.44	.13	.71	.65	.17	.19
Francisco de Vitoria	.32	.47	.50	.00	.00	.41	.04	.32	.00	.73	.54	.33	.43	.00	.00	.29	.00	.70	.72	.03	.05
1-Baseline (Aristotle)	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
Martha Nussbaum	.24	.34	.57	.00	.00	.42	.00	.27	.00	.59	.48	.00	.37	.00	.00	.39	.03	.57	.00	.10	.15
Johann Friedrich Herbart	.18	.00	.21	.12	.12	.03	.10	.06	.09	.08	.18	.18	.31	.21	.03	.18	.19	.18	.10	.04	.24
Friedrich Nietzsche	.01	.00	.00	.00	.00	.01	.00	.00	.00	.08	.03	.00	.00	.00	.00	.00	.00	.01	.04	.00	.00

Table 5: Achieved F_1-score on the main test dataset, from macro-precision and macro-recall ("All") and for each of the 20 value categories for the submission that achieved the highest "All" F_1-score per team. "Teams" in gray are shown for comparison: an ensemble using the best submission for each individual category (where such a best-in-a-category submission might not be in this table if a different submission of the same team reached a higher "All" F_1-score) and our BERT model and 1-Baseline.

Team	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Nahj al-Balagha																					
Best per category	.48	.18	.49	.50	.67	.66	.29	.33	.62	.51	.37	.55	.36	.27	.33	.41	.38	.33	.67	.20	.44
Adam Smith	.40	.13	.49	.40	.50	.65	.25	.00	.58	.50	.30	.51	.28	.24	.29	.33	.38	.26	.67	.00	.36
Tübingen (Stanley Grenz)	.35	.16	.39	.00	.36	.64	.25	.00	.30	.48	.26	.40	.28	.22	.22	.36	.25	.33	.67	.00	.44
MaChAmp (Robert S. Hartman)	.34	.08	.35	.40	.44	.61	.17	.00	.55	.40	.31	.33	.23	.24	.23	.30	.24	.25	.50	.07	.36
Hitachi (R. M. Hare)	.34	.08	.31	.17	.40	.62	.09	.33	.51	.49	.29	.45	.21	.14	.21	.28	.23	.27	.50	.00	.25
Mao Zedong	.32	.06	.39	.31	.44	.66	.10	.33	.59	.41	.16	.45	.24	.16	.31	.35	.20	.25	.25	.00	.28
Confucius	.31	.09	.31	.25	.36	.60	.12	.31	.57	.43	.19	.46	.22	.16	.26	.30	.18	.22	.29	.00	.24
Prodicus	.30	.17	.33	.00	.40	.59	.00	.00	.37	.42	.27	.53	.26	.07	.00	.38	.35	.23	.00	.17	.41
CSECU-DSG (Fazlur Rahman)	.29	.15	.29	.00	.40	.57	.00	.00	.00	.41	.24	.00	.25	.00	.09	.36	.38	.28	.67	.00	.28
SUTNLP (David Gauthier)	.29	.14	.22	.40	.00	.55	.00	.00	.33	.37	.24	.48	.27	.08	.23	.34	.30	.19	.20	.10	.34
BERT (Aristotle)	.28	.14	.09	.00	.67	.41	.00	.00	.28	.28	.23	.38	.18	.15	.17	.35	.22	.21	.00	.20	.35
Arthur Caplan	.28	.12	.26	.00	.27	.61	.13	.00	.55	.51	.20	.55	.24	.14	.23	.25	.19	.21	.33	.10	.29
Epicurus	.27	.09	.20	.00	.25	.49	.17	.12	.41	.38	.13	.47	.21	.27	.11	.29	.23	.24	.25	.06	.30
LRL_NC (George Boole)	.27	.07	.30	.29	.22	.55	.18	.00	.18	.45	.21	.29	.26	.27	.27	.29	.30	.21	.00	.10	.32
Noam Chomsky	.26	.09	.14	.00	.44	.54	.10	.13	.24	.50	.19	.42	.30	.13	.13	.34	.22	.21	.20	.11	.32
Aristoxenus	.25	.08	.13	.00	.29	.49	.00	.20	.21	.36	.23	.30	.27	.11	.00	.35	.24	.26	.18	.08	.32
Sina (Seyyed Hossein Nasr)	.25	.07	.21	.00	.40	.60	.12	.00	.12	.38	.19	.26	.22	.17	.22	.28	.18	.22	.29	.12	.27
Niccolò Machiavelli	.23	.12	.23	.00	.57	.47	.00	.00	.00	.45	.32	.22	.28	.15	.00	.29	.20	.22	.00	.00	.30
Augustine of Hippo	.23	.08	.16	.00	.10	.55	.09	.10	.39	.47	.14	.50	.23	.00	.10	.28	.23	.27	.08	.00	.27
Quintilian	.20	.03	.16	.00	.10	.46	.13	.20	.14	.38	.19	.49	.19	.00	.24	.25	.16	.18	.00	.00	.22
1-Baseline (Aristotle)	.13	.04	.09	.01	.03	.41	.04	.03	.23	.38	.06	.18	.13	.06	.13	.17	.12	.12	.01	.04	.14
The New York Times																					
Best per category	.47	.50	.22	-	.03	.54	.40	-	.50	.59	.52	-	.33	1.0	.57	.33	.40	.62	1.0	.03	.46
Hitachi (R. M. Hare)	.34	.22	.22	-	.00	.48	.40	-	.00	.53	.44	-	.18	1.0	.20	.12	.29	.55	.33	.00	.36
Mao Zedong	.32	.22	.12	-	.00	.47	.29	-	.22	.53	.41	-	.32	.50	.15	.21	.40	.56	.33	.00	.38
Arthur Caplan	.32	.14	.15	-	.00	.31	.00	-	.50	.59	.30	-	.18	1.0	.15	.11	.36	.40	.29	.00	.42
Confucius	.31	.18	.17	-	.00	.46	.20	-	.00	.55	.35	-	.18	1.0	.00	.13	.36	.46	.40	.00	.38
Adam Smith	.27	.33	.18	-	.00	.42	.00	-	.00	.58	.52	-	.18	.00	.00	.21	.31	.62	.50	.00	.46
Niccolò Machiavelli	.25	.50	.00	-	.00	.44	.00	-	.00	.53	.37	-	.00	.00	.00	.12	.23	.50	1.0	.00	.42
Sina (Seyyed Hossein Nasr)	.24	.11	.00	-	.00	.29	.00	-	.33	.57	.31	-	.23	.67	.00	.21	.31	.27	.33	.00	.38
BERT (Aristotle)	.24	.00	.00	-	.00	.29	.00	-	.00	.53	.43	-	.00	.00	.57	.26	.27	.36	.50	.00	.32
Augustine of Hippo	.19	.40	.10	-	.00	.34	.00	-	.00	.52	.26	-	.17	.00	.00	.19	.27	.23	.10	.00	.36
MaChAmp (Robert S. Hartman)	.19	.00	.20	-	.00	.21	.00	-	.00	.57	.40	-	.00	.00	.00	.14	.37	.33	.29	.00	.43
1-Baseline (Aristotle)	.15	.05	.03	-	.03	.28	.03	-	.05	.51	.20	-	.07	.03	.12	.12	.26	.24	.03	.03	.33
Quintilian	.10	.05	.03	-	.03	.28	.03	-	.00	.51	.00	-	.07	.00	.12	.12	.00	.00	.00	.03	.33

Table 6: Achieved F_1-score on the supplementary test datasets (Nahj al-Balagha and The New York Times), from macro-precision and macro-recall ("All") and for each of the 20 value categories for the submission that achieved the highest "All" F_1-score per team. "Teams" in gray are shown for comparison: an ensemble using the best submission for each individual category (where such a best-in-a-category submission might not be in this table if a different submission of the same team reached a higher "All" F_1-score) and our BERT model and 1-Baseline. The New York Times dataset contains no argument resorting to Stimulation, Power: resources, or Tradition.