Epicurus at SemEval-2023 Task 4: Improving Prediction of Human Values behind Arguments by Leveraging Their Definitions

Christian Fang ⋆
Department of Sociology
Utrecht University
c.fang@uu.nl

Qixiang Fang ⋆
Department of Methodology and Statistics
Utrecht University
q.fang@uu.nl

Dong Nguyen
Department of Information and Computing Sciences
Utrecht University
d.p.nguyen@uu.nl

Abstract

We describe our experiments for SemEval-2023 Task 4 on the identification of human values behind arguments (ValueEval). Because human values are subjective concepts which require precise definitions, we hypothesize that incorporating the definitions of human values (in the form of annotation instructions and validated survey items) during model training can yield better prediction performance. We explore this idea and show that our proposed models perform better than the challenge organizers' baselines, with improvements in macro F_1 scores of up to 18%.

1 Introduction

Human values are distinct beliefs that guide human behavior (Schwartz et al., 2012). Examples of such values are hedonism (i.e., seeking pleasure in life), face (i.e., maintaining recognition in society), and humility (i.e., being humble).

Studying human values has a long history in the social sciences and in studies of formal argumentation due to their manifold applications (Kiesel et al., 2022; Schwartz et al., 2012). For example, researchers might be interested in studying how the human values individuals subscribe to affect their charitable behavior (Sneddon et al., 2020) or voting behavior (Barnea and Schwartz, 1998). In NLP, human values can be leveraged for personality recognition (Maheshwari et al., 2017), or for assessing what values are implied in online discourses (Kiesel et al., 2022). To that end, Task 4 (ValueEval) of the SemEval 2023 competition (Kiesel et al., 2023) called for participants to design NLP systems that can classify a given argument as belonging to one of 20 value categories described in Kiesel et al.'s human value taxonomy (2022).

Human values are inherently subjective concepts, which becomes evident in, for example, the existence of many human value taxonomies, each of which contains somewhat different and differently-defined human values (e.g. Rokeach, 1973; Schwartz et al., 2012). Accordingly, in any scientific study of human values, clear definitions are key. Therefore, we argue that it is important to incorporate the definitions of human values into model training. Our approach leverages the definitions of human values (based on survey questions and annotation instructions), which we refer to as **definitional statements**, in a natural language inference (NLI) setup. This approach offers additional theoretical benefits, such as potentially higher model validity (i.e., more accurate encodings of human values) as well as greater prediction reliability (see §3). We showed that our approach achieved better performance than the challenge baselines. We also conducted additional post-hoc analyses to explore how prediction performance would vary by the number of definitional statements per value category. We found that even with only a few definitional statements per value category, our proposed approach achieved good performance. Our code is available in a public GitHub repository (https://github.com/fqixiang/SemEval23Task4).

2 Background

2.1 Task Setup

The goal of the challenge task was to, given a textual argument and a human value category, classify whether the argument entails that value category. Each argument consisted of a premise, stance (for or against), and conclusion and was assigned a binary label for each of the 20 human value categories (also called level-2 values in Kiesel et al. (2022)). Figure 1 illustrates this.

2.2 Related Work

Our approach of using definitions of human values in the form of annotation instructions and survey

⋆ Shared first authorship.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 221–229

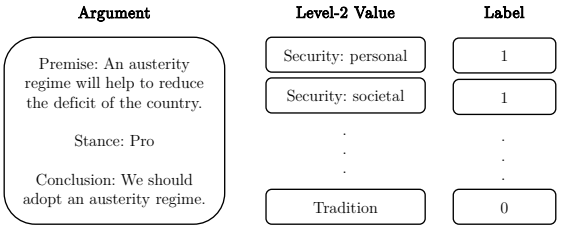

Figure 1: Illustration of the task.

questions is positioned within two streams of prior literature. The first stream used dictionary definitions and annotation instructions to improve, for instance, information extraction from legal documents (especially with a small data set) (Kang et al., 2021), retrieval of the semantic roles of tokens (Zhang et al., 2022), and detection of slang words (Wilson et al., 2020), dialects (Demszky et al., 2021), and rare words (Pilehvar and Collier, 2017). The second stream used survey questions for improving the prediction of social science constructs, such as personality traits (Kreuter et al., 2022; Vu et al., 2020; Yang et al., 2021), as well as social and political attitudes (Fang et al., 2022a).

3 A Measurement Problem

In the social sciences, detecting human values is a measurement problem and typically relies on the use of (validated) survey instruments, such as the Schwartz Value survey (Schwartz et al., 2012). Typically, respondents are asked to rate the importance of a given value (e.g., hedonism), presented in the form of survey questions, on a numerical scale. In this way, a numerical summary per human value can be assigned to each respondent.

Because human values are abstract concepts that are not directly observable, the respective measurements are likely to suffer from measurement error (Fang et al., 2022b). Social scientists are, therefore, particularly concerned about the content validity and reliability of such measurements. Content validity refers to an instrument fully capturing what it aims to measure (Trochim et al., 2016), whereas reliability means that measurements are stable over time and contexts (i.e., do not suffer from large random variations) (Trochim et al., 2016). To ensure high content validity, multiple survey questions that capture different sub-aspects of a value (e.g., hedonism) are typically used. Each respondent's answers to these questions about the same value are then aggregated (e.g., averaged) to obtain a single, more reliable score.

Likewise, incorporating definitional statements when training a model to predict human values from arguments might help to improve the content validity of the model, as well as the reliability of the predictions (Fang et al., 2022a). For instance, the human value "achievement" has many sub-aspects (i.e., being ambitions, having success, being capable, being intellectual, and being courageous). By incorporating these finer-grained definitions of "achievement" into model training, the model can learn to encode the full scope of this value, which can in turn help to identify arguments that entail this value. Furthermore, averaging a model's predictions across multiple definitional statements of the same human value category can lead to more reliable, less random results, which is consistent with the social sciences' approach.

4 System Overview

4.1 Data Augmentation with Definitional Statements

We created definitional statements for each of the 20 value categories based on two sources. The first source were the annotation instructions, which we obtained from the annotation interface provided by Kiesel et al. (2022). The second source were the survey questions that underlie the human value taxonomy uses in this challenge. We collected all relevant survey questions from the surveys that Kiesel et al. (2022) based their human value taxonomy on, namely the PVQ5X Value Survey (Schwartz et al., 2012); its predecessor, the Schwartz Value survey (Schwartz, 1992); the World Value Survey (Haerpfer et al., 2022); the Rokeach Value Survey (Rokeach, 1973); and the Life Values Inventory (Brown and Crace, 2002). For an overview of the number of definitional statements per value category, see Appendix A.

We harmonized all definitional statements by forcing them to adhere to a "It is important to be/have" sentence structure to prevent models from learning uninformative idiosyncratic formulations. Figure 2 shows such as an example.

Next, we augmented the training data set with definitional statements. We dropped "conclusion" and "stance" from the arguments, because per our observation of the data, the "premise"s alone already contain all the information about the underlying human values, which renders the use of "conclusion" and "stance" redundant. Each premise in

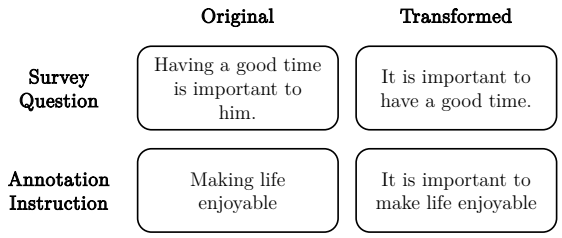

Figure 2: Example for original and transformed definitional statements for the value category "hedonism".

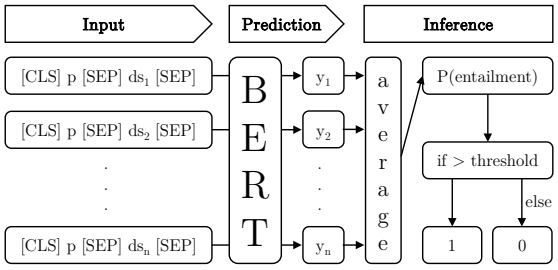

Figure 3: Illustration of the prediction step of the system for a given premise (p) and a value category described by n definitional statements (ds_i). We feed each combination of p and ds_i into a finetuned BERT model, obtain each individual binary prediction y_i, average these predictions to obtain the probability of entailment for a value category given the premise, and finally, make a binary decision based on a (finetuned) threshold.

the data set was combined with each definitional statement. For each combination of premise and definitional statement, we assigned "entailment" if the associated value label in the training data was 1, and "not entailment" otherwise.

4.2 NLI Setup

We used an NLI setup for modelling. NLI involves judging whether a hypothesis can be inferred from a premise. If so, then that premise entails the hypothesis. In our case, the textual premises from arguments constitute the premises, whereas the definitional statements constitute the hypotheses. We used BERT (Devlin et al., 2019) as the model of choice within the NLI setup.

4.3 Averaging Predictions and Thresholding

Our system yields a binary prediction for each combination of premise and definitional statement. Therefore, multiple predictions exist for every combination of premise and human value. To convert these multiple predictions into a single binary prediction per premise and value category, we averaged the predictions per value category, and applied a (fine-tuned) threshold to determine whether it is an entailment. Figure 3 illustrates this.

5 Experimental Setup

5.1 Data Splits

The main data set comprised 8,865 annotated arguments. We used the same split as the challenge organizers, namely a training set (61%), a validation set (21%), and a test set (18%). The label distribution of this data set was highly imbalanced: for example, only about 3.4% of all arguments were labelled "hedonism", whereas 47.6% were labelled "Universalism: concern". The second test set (Nahj al-Balagha) contained 279 annotated arguments from Islamic religious texts. The label distribution for that test set was slightly less imbalanced than the main data set (Mirzakhmedova et al., 2023).

5.2 Preprocessing and Hyperparameter Tuning

We used the pretrained "bert-base-uncased" and its tokenizer from Hugging Face (Wolf et al., 2020). To construct the input vector, we follow Jiang and de Marneffe (2019), where each premise and definitional statement is separated by the [SEP] token. We used a binary classification head to predict whether a given premise entailed a definitional statement, trained the model based on cross-entropy loss, and chose the best model with the lowest loss on the validation set.

We fine-tuned the number of training steps, with early stopping where the patience parameter was set to 10. We also tested ten different thresholds (ranging from 0 to 0.9 with increments of .1) on the validation set and chose the best-performing threshold for the test set based on macro F_1 scores on the validation set. The optimal threshold was 0 or 0.3, depending on the model. The full list of hyperparameters is in Appendix B.

5.3 Trained Models

We fine-tuned four BERT models (see Table 1) based on the type of definitional statements (annotation instructions or survey questions) and loss functions (weighted or unweighted). Weighted cross-entropy loss is considered to account for the imbalanced class distribution in the training data. The weights were calculated as proportional to inverse class distributions in a training batch.

Model	Definitional statements	Weighted loss?	Training size
ANN_{uw}	annotations		31,807,742
ANN_{w}	annotations	✓	31,807,742
SVY_{uw}	survey items		37,109,033
SVY_{w}	survey items	✓	37,109,033

Table 1: Overview of trained models

5.4 Evaluation Metrics

We evaluated model performance by the macro F_1 score, calculated as the average of all 20 individual F_1 scores. During model training and validation, we were not aware that the challenge organizers used a different method for calculating macro F_1, namely by using the averages of precision and recall. Therefore, to stay consistent with our training and validation strategy, we focus on discussing the results based on our macro F_1 calculation. This calculation method has also been shown to be the more appropriate metric between the two, especially for model rankings and when data is imbalanced (Opitz and Burst, 2019). However, for completeness, we show both results and discuss their differences.

6 Results

Table 1 shows the results of our four model runs on the main testing data set (Main), as well as the testing set comprising arguments from religious texts (Nahj al-Balagha).

6.1 Main Test Set

On the main test set, all four models performed better than the two baselines. The best-performing model (ANN_{w}) used annotations and weighted cross-entropy loss, achieving an F_1 score of .45 (15% higher than the BERT baseline). The worst performing model (ANN_{uw}) still achieved a 10% increase in macro F_1 over the BERT baseline. Compared to the BERT baseline, the best performing model achieved substantially higher F_1 scores for stimulation (.13 vs .05), face (.29 vs .13) and humility (.21 vs .07), whereas prediction performance was worse for, amongst others, hedonism (.15 vs .20), security: personal (.70 vs .74), and conformity: interpersonal (.23 vs .19). Overall, on the main test set, the weighted models performed better than the unweighted models, and the models with annotation instructions performed better than those with survey questions. This is unsurprising, as using the annotation instructions probably more directly captures how the annotators came to label the data, whereas survey items are more distal.

6.2 Religious Texts Test Set

Since the models were not trained on this data set, good prediction performance requires that the trained models generalize well to texts from a (very) different distribution. Three out of the four trained models performed better than the BERT baseline, and one model performed equally well. The best-performing model used survey questions and an unweighted loss function (SVY_{uw}) and achieved an 18% higher F_1 score than the BERT baseline. The pattern of model performance is different than on the main data set. Specifically, the models including survey items performed better than the ones including annotation instructions, which might indicate that using survey items (which are more distal measures of human values than annotation instructions) may help especially when predicting out-of-distribution arguments. The unweighted models performed better than the weighted ones, which is surprising. The best model achieved substantially higher F_1 scores compared to the BERT baseline for, amongst others, power: resources (.25 vs .00), face (.52 vs .28) and universalism: nature (.33 vs .00), and worse scores for universalism: tolerance (.00 vs .20) and hedonism (.40 vs .67).

Note that if we abide by the challenge organizers' macro F_1 calculation, the ranking of the models relative to each other and to the BERT baselines can be different. Especially on the test set comprising religious texts, per the organizers' calculation, none of our models outperformed the BERT baseline, and two models out of the four models achieved the same macro F_1 score as the BERT baseline.

6.3 Influence of the Number of Definitional Statements on Macro F_1

While our models achieved higher prediction performances than the challenge owners' baselines, a limitation of our approach is that, even with a relatively small number of training arguments/premises (<6,000), the total number of training instances can be very large, as this also depends on the number of value categories and definitional statements. In our experiments, one model took about 20 GPU hours to train. Computing times might become impractically long when the number of arguments, values and definitional statements increases.

Test set / Approach	macro F_1 (our own)	macro F_1 (official)	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
Main																						
Best per category	.588	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.551	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.391	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.249	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
ANN_{uw}*	.431	.44	.47	.59	.13	.15	.57	.33	.50	.29	.70	.59	.47	.54	.19	.21	.50	.19	.69	.72	.33	.45
ANN_{w}	.450	.46	.49	.59	.22	.33	.57	.36	.50	.23	.70	.61	.47	.45	.26	.19	.47	.28	.68	.74	.34	.52
SVY_{uw}*	.434	.45	.46	.56	.18	.31	.58	.35	.55	.20	.70	.58	.44	.50	.11	.21	.48	.26	.69	.75	.34	.45
SVY_{w}	.435	.44	.47	.58	.21	.22	.56	.32	.48	.26	.70	.59	.41	.47	.22	.14	.48	.27	.69	.72	.37	.53
Nahj al-Balagha																						
Best per category	.428	.48	.18	.49	.50	.67	.66	.29	.33	.62	.51	.37	.55	.36	.27	.33	.41	.38	.33	.67	.20	.44
Best approach	.40	.356	.13	.49	.40	.50	.65	.25	.00	.58	.50	.30	.51	.28	.24	.29	.33	.38	.26	.67	.00	.36
BERT	.2155	.28	.14	.09	.00	.67	.41	.00	.00	.28	.28	.23	.38	.18	.15	.17	.35	.22	.21	.00	.20	.35
1-Baseline	.121	.13	.04	.09	.01	.03	.41	.04	.03	.23	.38	.06	.18	.13	.06	.13	.17	.12	.12	.01	.04	.14
ANN_{uw}*	.252	.28	.10	.22	.00	.18	.52	.11	.00	.55	.40	.25	.54	.26	.24	.24	.30	.29	.25	.25	.05	.28
ANN_{w}	.2155	.24	.16	.17	.00	.18	.47	.08	.12	.46	.37	.31	.39	.15	.06	.15	.31	.23	.19	.13	.06	.32
SVY_{uw}*	.254	.28	.10	.24	.00	.40	.50	.09	.25	.52	.41	.24	.44	.19	.10	.25	.27	.27	.19	.33	.00	.28
SVY_{w}	.231	.26	.18	.20	.00	.17	.52	.04	.12	.50	.40	.22	.49	.19	.10	.24	.30	.25	.25	.12	.04	.29

Table 2: Achieved F_1-score of team Epicurus per test dataset (macro and for each of the 20 value categories). Our own macro F_1 is the unweighted average of the 20 individual F_1 scores, while the official macro F_1 is calculated using the averages of precision and recall over all 20 value categories. Approaches marked with * were not part of the official evaluation. Approaches in gray are shown for comparison: an ensemble using the best participant approach for each individual category, the best participant approach, and the organizer's BERT and 1-Baseline.

Therefore, to investigate the scalability of our proposed approach, we explored the influence on the number of definitional statements per value category on our approach's performance. Note that because these analyses were conducted after the challenge's submission deadline, their results were not part of the official submissions. Additionally, in view of limited computational resources, we limited our additional analyses to the ANN_{w} model. We tested ten different sample sizes – ranging from one to ten definitional statements per value category – with the respective number of definitional statements sampled using simple random sampling with replacements (to circumvent the issue of some value categories having a smaller number of definitional statements than the sample size of interest).

On the main test set, the highest macro F_1 was obtained with two definitional statements per value category (own macro F_1: 0.458; challenge organizer's F_1: 0.472; see 4). Macro F_1 decreased substantially as the number of definitional statements increased to four, and levelled off after that. We observed this trend for both macro F_1 calculation methods. On the test set comprising arguments from religious text, macro F_1 scores varied across the number of definitional statements. The best performance was obtained for a sample size of eight, while at sample sizes between two and four the achieved F_1 scores were already higher than the original ANN_{w} model.

These results show that even with just two or three definitional statements per value category, our proposed approach could achieve higher or comparable performance than when all available definitional statements are used, while the computational overhead is reduced by about 90 per cent. This suggests that our proposed approach is scalable by reducing the number of definitional statements per value category.

For these additional analyses, we expected the models' performance to increase with larger sample sizes, before eventually levelling off. However, the observed pattern was that, after peaking, performance decreased and then levelled off. A potential

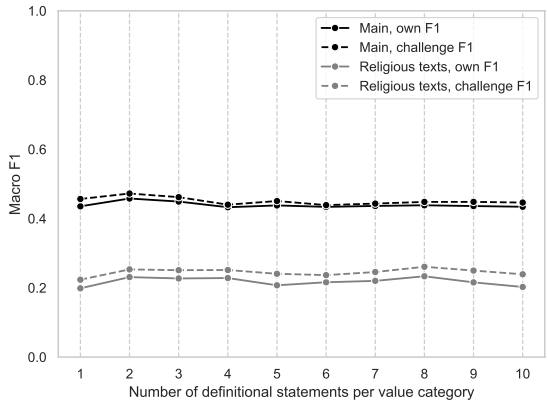

Figure 4: Achieved macro F_1 on the main test set and religious texts test set, per number of definitional statements per value category.

reason might be that most model hyperparameters (e.g., batch size, learning rate) were the same for all studied sample sizes (except for, for instance, step size, which depends on the training size), where these specific hyperparameter values might be potentially inappropriate for some of the models (especially those with substantially training sizes) to efficiently learn from the data.

7 Conclusion and Discussion

Our models achieved higher prediction performances than the challenge owners' baselines, indicating that there is merit to using definitional statements (i.e., annotation instructions and survey questions) for predicting the human values implied in textual arguments. This aligns with previous studies that incorporated dictionary definitions or survey questions for better task performance (e.g. Kreuter et al., 2022; Fang et al., 2022a). Furthermore, we showed that by using just a small number of definitional statements per value category, we could achieve prediction performance comparable to (and in some cases, better than) when all available definitional statements are used, while significantly reducing computational overhead.

As the goal of our study was not to obtain the best performance possible, but to test the idea that incorporating definitional statements into model training would improve prediction of human values, we did not try more advanced or larger language models, or techniques that could have improved prediction performance, such as paraphrasing (Wei and Zou, 2019) and ensemble learning. We also fine-tuned our models on only a limited set of hyperparameters.

A reason for why even the best team's model achieved a macro F_1 score of only .551 could be the low inter-rater agreement of the annotations. The average Krippendorff's α was just 0.49 for level-1 value categories (i.e., sub-values of level-2 categories) (Kiesel et al., 2022). To investigate the inter-rater agreement for the level-2 values, which are the focus of this challenge, two of the authors annotated a random sample of 100 arguments from the training data and arrived at an even lower α of just 0.31. Accordingly to Krippendorff (2004, p. 241), α values below 0.667 reflect very poor inter-rater agreement and high random (measurement) error. A reason for this low inter-rater agreement might be the annotation scheme requiring annotators to classify arguments as relating to 54 values — each of which with several instructions — which can overwhelm even experienced annotators. Furthermore, some values and their associated explanations seem similar, such as "Power: resources" — which partially concerns having wealth — and "Security: personal" — which partially concerns not having debts and having a comfortable life. Therefore, classifying an argument as belonging to a particular value may be more subjective than intended. Improvements in the human value taxonomy and/or the annotation scheme are likely needed to yield more reliable and valid measurements of human values.

Acknowledgement

This work was partially supported by the Dutch Research Council (grant number VI.Vidi.195.152 to D. L. Oberski).

References

Marina F Barnea and Shalom H Schwartz. 1998. Values and voting. *Political Psychology*, 19(1):17–40.

Duane Brown and R Kelly Crace. 2002. Life values inventory: Facilitator's guide. *Williamsburg, VA*.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qixiang Fang, Anastasia Giachanou, and Ayoub Bagheri. 2022a. Modelling stance detection as textual entailment recognition and leveraging measurement knowledge from social sciences. *arXiv preprint arXiv:2212.06543*.

Qixiang Fang, Dong Nguyen, and Daniel L. Oberski. 2022b. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, 11:1–31.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey: Round seven—country-pooled datafile version 4.0. *JD Systems Institute: Madrid, Spain*.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating bert for natural language inference: A case study on the commitmentbank. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6086–6091.

Liangyi Kang, Jie Liu, Lingqiao Liu, and Dan Ye. 2021. Label definitions augmented interaction model for legal charge prediction. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 270–283. Springer.

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*, Toronto, Canada. Association for Computational Linguistics.

Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. Items from psychometric tests as training data for personality profiling models of Twitter users. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 315–323, Dublin, Ireland. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content analysis*. SAGE Publ., Thousand Oaks, CA.

Tushar Maheshwari, Aishwarya N Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. 2017. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 731–741.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, 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.

Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

The pandas development team. 2022. pandas-dev/pandas: Pandas.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mohammad Taher Pilehvar and Nigel Collier. 2017. Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 388–393, Valencia, Spain. Association for Computational Linguistics.

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

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

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):663.

Joanne N Sneddon, Uwana Evers, and Julie A Lee. 2020. Personal values and choice of charitable cause: An exploration of donors' giving behavior. *Nonprofit and Voluntary Sector Quarterly*, 49(4):803–826.

William MK Trochim, James P Donnelly, and Kanika Arora. 2016. *Research methods: the essential knowledge base*. Cengage Learning, Boston, MA.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1512–1524.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang nlp applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142.

Li Zhang, Ishan Jindal, and Yunyao Li. 2022. Label definitions improve semantic role labeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5613–5620.

A Overview of Definitional Statements, Per Value Category

See next page.

B Other Implementation Details

Computing Infrastructure All analyses were done on one of the High Performance Computing (HPC) cluster offered by Utrecht University. Python 3.10, PyTorch 1.12.1 (Paszke et al., 2019), torchtext 0.13.1, Huggingface Transformers 4.24.0 (Wolf et al., 2020) and CUDA 11.3 were used for finetuning BERT (Devlin et al., 2019) and predicting the labels of the test set. Numpy 1.13.1 (Harris et al., 2020), pandas 1.4.4 (pandas development team, 2022), and scikit_learn 1.1.3 (Pedregosa et al., 2011) were used for data wrangling.

Runtime About 20 hours per model on an RTX 6000 GPU node.

Number of parameters 110 million.

Validation performance ANN_{uw}: 0.432; ANN_{w}: 0.441; SVY_{uw}: 0.423; SVY_{w}: 0.434.

Hyperparamters For the BERT models, we used the following hyperparameters:

- num_train_epochs=5
- per_device_train_batch_size=128
- per_device_eval_batch_size=1024
- warmup_steps=250
- weight_decay=0.01
- early stopping criterion: step
- patience for early stopping: 10

	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**	**Total**
Annotation instructions	18	17	15	6	26	11	7	9	28	12	12	13	8	12	28	11	18	18	12	13	294
Survey questions	18	19	15	24	31	11	9	13	28	14	21	20	13	16	30	14	17	14	9	7	343
Total	36	36	30	30	57	22	16	22	56	26	33	33	21	28	58	25	35	32	21	20	637

Table 3: Overview of the number of annotation instructions, survey questions, and sum of the two, per value category.