Shared Task 4 - ValueEval: Identification of Human Values behind Arguments

Muskan Goyal, Nikhil Hulle, Rithik Kumar
University of Colorado Boulder
{muskan.goyal, nikhil.hulle, rithik.athiganursentil}@colorado.edu

Abstract

This paper examines the human values behind various arguments using Natural Language Processing. Values are commonly accepted answers to why some option is desirable in the ethical sense and are thus essential both in real-world argumentation and theoretical argumentation frameworks. However, their large variety has been a major obstacle to modelling them in argument mining. Therefore to address this issue, we use a dataset with 54 values and 5270 arguments to perform the automatic classification of human values. We achieve promising results and we support our findings by performing extensive result analysis.

1 Introduction

People can have disagreements about opinions on many controversial issues. Asking them repeatedly why they consider something desirable is a good technique to get to the root of such disagreements. People differ in their priorities and opinions about what is generally worthwhile to strive towards and how to accomplish so, which are sometimes referred to as human values (Searle, 2001). Due to their importance, human values have been researched for decades in the social sciences (Schwartz, 1994)as well as in formal reasoning (Bench-Capon, 2003).

Human values thus provide the context for classifying, contrasting, and evaluating argumentative statements in computational linguistics, opening up a number of possibilities, including the ability to generate or select arguments based on the value system of a target audience and to identify opposing and shared values on both sides of a contentious issue.

Due to their abundance, frequent implicit use in arguments, and hazy definitions, the process of identifying values in arguments appears overwhelming. However, the expansion of argumentation datasets, improvements in Natural Language Understanding, and the careful taxonomization of values by social scientists over a decade have made such an automatic identification feasible (Kiesel and Alshomary, 2022).

For this study, we use a dataset (defined in (Kiesel and Alshomary, 2022)) that has a multi-level taxonomy of 54 human values taken from four authoritative cross-cultural social science studies, with a total of 5270 arguments from the US, Africa, China, and India, each of which are manually annotated for all values, corresponding to about 850k human judgments.

In this paper, we perform automatic identification of human values in written arguments. Classification of human values is performed for the all the taxonomy levels using various models. We implement, train and evaluate multiple models to perform comparative analysis. Results are reported per taxonomy level showing positive outcomes both inside and across cultures.

2 Background

The majority of social sciences, if not all of them, are concerned with human values, and computational frameworks of argumentation (Bench-Capon, 2003) have been incorporated into this concern. Values have not yet been examined in NLP for argument mining, as was done in this study, but they have been for personality profiling (Das et al., 2017).

2.1 Human Values in Social Science

(Boehm, 1974) estimates the total number of human values to be less than hundreds and creates a practical survey of 36 values that distinguishes between values pertaining to desirable end states and desirable behaviour (Boehm, 1974). He does this by combining research from anthropology, sociology, philosophy, and psychology.

48 value questions were generated by (Schwartz et al., 2012) specifically for the cross-cultural study of the universal needs of people and communities, such as following the law and being modest. Additionally, (Schwartz, 1994) suggests that values are related by their propensity to be compatible in their pursuit. This relatedness represents two ”higher order” conflicts that allow for the analysis of values at several levels: (1) openness to change/own thinking vs. conservation/submission, and (2) self-transcension (directed towards others/the environment) vs. self-enhancing (focused towards one's self).

A ”meta-inventory” of 16 values, including honesty and justice, which (Cheng and Fleischmann, 2010) combine 12 schemes into, revealed a significant amount of overlap in schemes across academic domains. We do not, however, further explore the meta-inventory in this study because it is strictly more coarse-grained than Schwartz et al theory.

2.2 Human Values in Argumentation Research

Formal argumentation uses value systems to simulate audience-specific preferences; as a result, the persuasiveness of an argument depends on how highly the audience holds the values it invokes. Examples include defeasible logic programming (Teze et al., 2019), BenchCapon's value-based argumentation framework (Bench-Capon, 2003), and value-based argumentation techniques (van der Weide et al., 2009). The latter is a development of the Dung framework (Dung, 1995) for abstract argumentation that has already been manually used to examine interactions between reasoning and persuasion subject to a certain value system (Atkinson and Bench-Capon, 2021).

3 System Overview

In this section, we describe all the models that we used for comparative analysis. We use Hugging-Face Transformers (Wolf et al., 2021), PyTorch (Paszke et al., 2019), PyTorchLightning (Falcon et al., 2020) and Scikit-learn (Pedregosa et al., 2011) for model implementation, training and evaluation.

3.1 1-Baseline

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

3.2 SVM

Finding a hyperplane in an N-dimensional space (N is the number of features) that categorizes the data points clearly is the goal of the support vector machine algorithm (Joachims, 1998). There are a variety of different hyperplanes that might be used to split the two classes of data points. Finding a plane with the greatest margin—that is, the greatest separation between data points from both classes—is our goal. Maximizing the margin distance adds some support, increasing the confidence with which future data points can be categorised.

Decision boundaries known as hyperplanes assist in categorizing the data points. Different classes can be given to the data points that fall on each side of the hyperplane. Additionally, the amount of features affects how big the hyperplane is. The hyperplane is essentially a line if there are just two input features. The hyperplane turns into a two-dimensional plane if there are three input features. When there are more than three features, it gets harder to imagine.

3.3 BERT

The primary technological advancement of BERT (Devlin et al., 2019) is the application of Transformer's bidirectional training, a well-liked attention model for language modelling. In contrast, earlier research looked at text sequences from either a left-to-right or a combined left-to-right and right-to-left training perspective. The study's findings demonstrated that bidirectionally trained language models can comprehend the context and flow of language more deeply than single-direction language models. The authors of the paper described a unique method called Masked LM, which made bidirectional training possible in models where it was previously not practicable.

The **Bert tiny, small and medium** are the pre-training compact models, in NLP large and expensive models that utilize an easy amount of general-domain text through self-supervised pre-training. Due to the cost of this, several model compression techniques on pre-trained language representation have been proposed.

3.4 BERT-tiny

This is one of the smaller and compact pre-trained BERT variants with 2 layers and 128 hidden units.

3.5 BERT-mini

This too is a smaller pre-trained BERT variant with 4 layers and 256 hidden units.

3.6 BERT-medium

This pre-trained BERT variant has total of 8 layers and 512 hidden units.

3.7 RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Accuracy) model (Liu et al., 2019) is a more robust version of BERT that is trained with a lot more data. It is based on fine-tuning the hyper-parameters that has improved the results and performance of the model significantly. To boost performance of BERT, RoBERTa also modified its training procedure and architecture. These modifications include removing next sentence prediction and dynamically changing the masking pattern during pre-training.

3.8 DistilBERT

DistilBERT (Sanh et al., 2019) is a BERT-based Transformer model that is compact, quick, affordable, and light. In order to shrink a BERT model by 40% during the pre-training stage, knowledge distillation is used. The authors offer a triple loss that combines language modeling, distillation, and cosine-distance losses in order to take advantage of the inductive biases that larger models acquire during pre-training.

4 Experimental Setup

4.1 Dataset

This dataset (Kiesel and Alshomary, 2022) has total of 54 human values and 5270 arguments with 5 levels of taxonomy: level 1 has 54 human values, level 2 has 20 human values, level 3 has 4 values, level 4a has 2 values, and level 4b has 4 values. In this paper, we perform classification for each level.

The dataset is also composed of 4 parts: *Africa, China, India, and USA*. For *Africa*, there are 50 arguments that were manually extracted from recent editorials of the debating ideas section of a pan-African news platform, *African Arguments*. 100 arguments are taken for *China* from the recommendation and hotlist section of a Chinese question-answering website, *Zhihu*. Similarly, *India* also has 100 arguments extracted from a famous blog i.e. the controversial debate topics 2021 section of *Group Discussion Ideas*. The largest number of arguments were collected for the *USA*. It had a total of 5020 arguments with a manual argument quality rating of at least 0.5 from the 30,497 arguments of the IBM-ArgQ-Rank-30kArgs dataset (Kiesel and Alshomary, 2022).

The number of arguments in other classes (250) is minimal when compared to the USA part, therefore, these other classes are mainly used for testing the robustness of the models in identifying values behind arguments. For each part/culture, the dataset is divided into 3 parts i.e. train, test, and validation. 85% arguments are used for training, 5% for validation, and rest 10% for testing.

The conclusions were chosen so that the various sets nearly contain the required number of arguments. Unfortunately, the different sets' value distributions varied as a result of this process. For our study, we thought the conclusion-wise split was more crucial because we want to discover if classifiers can generalize to conclusions that haven't yet been observed.

4.2 Models Used

This section shows an attempt at using conventional methods to automatically determine human values. This experiment focuses on the classification of human values behind different arguments in cross-cultural settings (USA, Africa, India, and China). We use various models and approaches for which we provide the implementation. The implementation can be found on our GitHub repo.

https://github.com/hulle123/SharedTask

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

SVM A linear kernel scikit-learn support vector machine trained label-wise with C = 18, 20, 22 and 24, each with max iterations 10000. C is a regularization parameter that controls the trade-off between achieving a low training error and a low testing error i.e the ability to generalize your classifier to unseen data.

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

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

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

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

Model	Level 1				Level 2				Level 3				Level 4a				Level 4b			
	Afi.	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**
1-Baseline	0.16	0.13	0.12	0.16	0.27	0.23	0.21	0.3	0.63	0.65	0.62	0.75	0.8	0.88	0.79	0.92	0.92	0.91	0.9	0.96
BERT	0.20	0.21	0.25	0.38	0.37	0.41	0.34	0.60	0.68	0.71	0.71	0.82	0.88	0.81	0.92	0.92	0.92	0.92	0.90	0.96
BERT-tiny	0.12	0.16	0.11	0.15	0.20	0.21	0.24	0.25	0.67	0.63	0.62	0.68	0.92	0.81	0.87	0.83	0.96	0.92	0.91	0.81
BERT-small	0.13	0.16	0.13	0.16	0.28	0.28	0.25	0.27	0.71	0.62	0.59	0.66	0.92	0.81	0.87	0.82	0.96	0.92	0.91	0.90
BERT-medium	0.13	0.16	0.16	0.16	0.26	0.29	0.26	0.39	0.71	0.63	0.62	0.68	0.92	0.8	0.88	0.83	0.96	0.92	0.92	0.9
RoBERTa	0.22	0.21	0.22	0.25	0.32	0.36	0.32	0.33	0.71	0.66	0.65	0.69	0.92	0.80	0.89	0.82	0.95	0.92	0.92	0.90
DistilBERT	0.19	0.19	0.19	0.26	0.29	0.32	0.28	0.33	0.70	0.67	0.60	0.68	0.92	0.82	0.87	0.82	0.95	0.92	0.91	0.90

Table 1: Macro F1-scores on each test set over all labels by level using various models

Model	Level 1				Level 2				Level 3				Level 4a				Level 4b			
	Afi.	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**	**Afi.**	**Chi.**	**Ind.**	**USA**
1-Baseline	0.08	0.07	0.06	0.08	0.16	0.13	0.11	0.18	0.46	10.48	0.44	0.6	0.66	0.78	0.66	0.85	0.85	0.84	0.81	0.92
BERT	0.93	0.92	0.92	0.94	0.85	0.86	0.87	0.91	0.74	0.73	0.61	0.71	0.86	0.76	0.80	0.74	0.92	0.87	0.85	0.84
BERT-tiny	0.92	0.92	0.92	0.94	0.82	0.85	0.87	0.89	0.70	0.73	0.57	0.70	0.85	0.73	0.78	0.74	0.92	0.85	0.84	0.81
BERT-small	0.92	0.91	0.93	0.94	0.84	0.85	0.87	0.88	0.73	0.71	0.53	0.71	0.85	0.74	0.77	0.73	0.92	0.85	0.84	0.81
BERT-medium	0.92	0.91	0.92	0.94	0.84	0.85	0.86	0.90	0.73	0.72	0.56	0.70	0.85	0.71	0.78	0.73	0.92	0.85	0.84	0.81
RoBERTa	0, 92	0.90	0.93	0.94	0.85	0.84	0.75	0.89	0.73	0.72	0.57	0.7	0.86	0.74	0.80	0.74	0.91	0.85	0.84	0.81
DistilBERT	0.92	0.90	0.92	0.94	0.84	0.85	0.86	0.90	0.7	0.75	0.60	0.71	0.86	0.75	0.78	0.73	0.91	0.85	0.84	0.82

Table 2: Accuracy on each test set over all labels by level using various models

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

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

4.3 Evaluation

The label-wise F1-score and its mean across all labels (macro-average), as well as its component precision and recall, are the main subjects of our evaluation. We provide accuracy for completeness, however, the skewed label distribution makes it less than ideal. For all metrics in the evaluation, macro-averages are used to give each value the same weight. It should be noted that the 1-Baseline is particularly effective for the F1-score because it always obtains a recall of 1. By definition, this baseline achieves F1-scores that are at least as high—and frequently higher—than label-wise random guessing based on label frequency.

5 Results

Here, we use the same techniques across all test sets to evaluate classification resilience without retraining. Due to the size of the non-US components, 28% of the values are devoid of supporting information. **Table 1** presents the Macro F1 scores and **Table 2** reports the accuracy of all the models that we experimented with. All the models had a batch size of 8, a learning rate of 2^{-5} and 20 epochs. **Table 3** presents the Macro F1 scores of the SVM model with different C values and max_iterations of 10000.(See Section 4.2). All the prediction files are present on our GitHub repo.

https://github.com/hulle123/SharedTask

6 Conclusion

It is difficult yet vital to computationally identify the human values that underlie arguments. This paper uses a dataset with a multilevel taxonomy of 54 values and 5270 arguments from four sources and conducts empirical analyses that compare various cultures and span many value granularity levels.

Even while more information and research seem to be required to reach this conclusion, the findings of this report provide evidence that adopting a cross-cultural value taxonomy could produce reliable ways of determining the values behind arguments. The logical next step based on this work is to conduct analyses that fully take advantage of label associations. Here, hierarchical classification methods (Babbar et al., 2013) show promise; multi-label categorization earning rules (Mencía and Janssen, 2016)can also reveal information about value linkages.

Values have a significant role in the strength of an argument, and extensive web data mining could enhance all phases of argument formulation, evaluation, and classification (Bench-Capon, 2021). For instance, comparing the values of opposing and supportive arguments can be useful. Misunderstandings between people and automated argumentation systems could be avoided by stating the val-

Country	C=18					C=20					C=22					C=24				
	1	2	3	4a	4b	1	2	3	4a	4b	1	2	3	4a	4b	1	2	3	4a	4b
Africa	0.15	0.26	0.63	0.84	0.87	0.16	0.19	0.56	0.78	0.85	0.20	0.19	0.66	0.78	0.85	0.16	0.19	0.55	0.78	0.85
China	0.12	0.24	0.53	0.77	0.84	0.12	0.24	0.53	0.77	0.85	0.12	0.23	0.52	0.78	0.84	0.12	0.25	0.56	0.74	0.82
India	0.17	0.23	0.56	0.71	0.81	0.18	0.23	0.56	0.71	0.81	0.18	0.24	0.57	0.73	0.81	0.18	0.24	0.56	0.73	0.82
USA	0.16	0.26	0.63	0.84	0.87	0.15	0.26	0.63	0.84	0.86	0.17	0.25	0.63	0.84	0.87	0.15	0.26	0.63	0.84	0.87

Table 3: Macro F1-scores on each test set over all labels by level using **SVM** with different C values and **Max-iterations as 10000**

ues behind arguments in clear terms (Kiesel and Alshomary, 2022). Similarly to that, an ”objective” examination of the shared principles that underlie divisions between political parties could be a start toward settling disputes that seem to have very deep roots.

Finally, social science researchers may be interested in the examination of values in huge text corpora. What values are communicated online? One could even track references to values over time using Internet archive data. We, therefore, expect that this work can act as a starting point for further research into how the general public perceives and experiences human values in modern (digital) life.

References

Katie Atkinson and Trevor J. M. Bench-Capon. 2021. Value-based argumentation. FLAP, 8:1543–1588.

Rohit Babbar, Ioannis Partalas, Éric Gaussier, and Massih-Reza Amini. 2013. On flat versus hierarchical classification in large-scale taxonomies. In NIPS.

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

Trevor J. M. Bench-Capon. 2021. Audiences and argument strength.

Werner W. Boehm. 1974. The nature of human values. by milton rokeach. new york: Free press, 1973. 438 pp. $13.95. Social Work, 19:758–759.

An-Shou Cheng and Kenneth R. Fleischmann. 2010. Developing a meta-inventory of human values. In ASIS&T Annual Meeting.

Amitava Das, Björn Gambäck, Tushar Maheshwari, Aishwarya N. Reganti, Samiksha Gupta, Anupam Jamatia, and Upendra Kumar. 2017. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In EACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell., 77:321–358.

William Falcon, Jiří Borovec, Nicholas Scott Eggert, Vadim Bereznyuk, Ir dXD, Adrian Wälchli, Jeremy Jordan, Sebastian Kazmarek Præsius, Tullie Murrell, Ethan Harris, Shreya V. Bapat, Hendrik Schröter, Akshay Kulkarni, Verena Haunschmid, Dmitry Lipin, Alok Pratap Singh, Thomas J. Fan, Nicki Skafte, Hadrien Mary, Cristobal Eyzaguirre, cinjon, Anton Bakhtin, Zhai Zh, Yongrae Jo, Peter Izsak, Oscar A. Rangel, Jeffrey Ling, Harsh Sharma, Elliot Waite, and Ayberk Aydin. 2020. Pytorchlightning/pytorch-lightning: Tpu support & profiling.

Thorsten Joachims. 1998. Making large scale svm learning practical. Technical reports.

Johannes Kiesel and Milad Alshomary. 2022. Identifying the human values behind arguments. In ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692.

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

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. In Neural Information Processing Systems.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Ron J. Weiss, J. Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12:2825–2830.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.

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, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. Journal of personality and social psychology, 103 4:663–88.

John R. Searle. 2001. Rationality in action.

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, 87:293 – 319.

Tom van der Weide, Frank Dignum, John-Jules Ch. Meyer, Henry Prakken, and Gerard A. W. Vreeswijk. 2009. Practical reasoning using values. In ArgMAS.

Thomas Wolf, Lysandre Debut, Sylvain Gugger, Patrick von Platen, Julien Chaumond, Stas Bekman, Sam Shleifer, Victor Sanh, Manuel Romero, Funtowicz Morgan, Suraj Patil, Julien Plu, Aymeric Augustin, Rémi Louf, Stefan Schweter, Denis, Matt, Nicolas Patry, erenup, Joe Davison, Kevin Canwen Xu, Philippe Schmid, Teven, Anthony Moi, Piero Molino, Grégory Châtel, Bram Vanroy, Philip May, Clément., and Daniel Stancl. 2021. huggingface/transformers: v4.9.2: Patch release.