Identifying the Human Values behind Arguments

Vatsal Verma
University of Colorado, Boulder
vatsal.verma@colorado.edu

Rohit Taware
University of Colorado, Boulder
rohit.taware@colorado.edu

Abstract

This paper studies human values which are mostly implicit and underlie in natural language arguments, such as the Concern under self-transcendence or dominance under self-enhancement. Values are generally agreed-upon explanations of why a particular alternative is preferable in an ethical sense, and they are therefore crucial to both practical argumentation and theoretical argumentation frameworks. However, modeling them in argument mining has been significantly hampered because of wide variations. To get over this problem, we will be using a multi-level taxonomy of 54 values that operationalizes human values. Additionally, we trained the data on a dataset containing 5270 arguments from four different geographic cultures named the USA, China, India, and Africa and this data has been manually annotated for moral principles.

1 Introduction

Why do people have different opinions on the best way to proceed in contentious situations, even though they use the same facts to do so? In order to find the answer to such a question we can ask people repeatedly what makes something desirable for them and the answer for the same can be given as human values[1]. Figure 1[2] illustrates how some deals tend to conflict while others align, which can lead to disputes over the right course of action but also to support. Choosing and prioritizing one value above other could be mainly because of differences between cultures and disagreement thereon. Various papers studied social sciences[3] and formal argumentation[4].

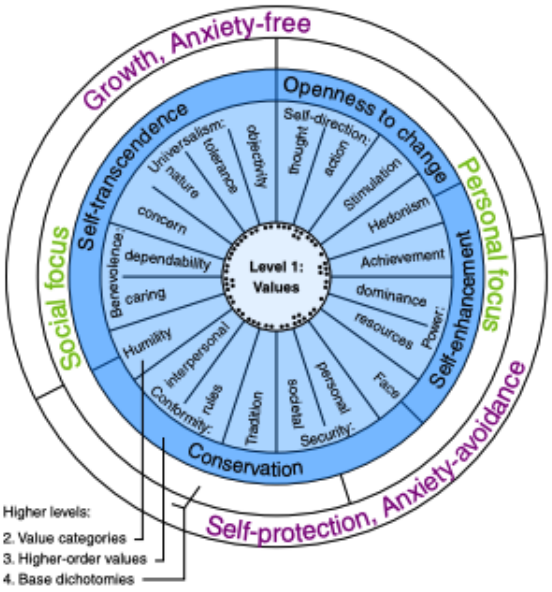

Figure 1[2]

According to various research in social sciences, a value is made up of [5][3] (1) belief (2) pertaining to desirable end states or modes of conduct which transcends specific situations, guides selections or evaluation of behavior, people and events and is ranked by importance relative to other values for forming the system of value priorities. Let's consider an example,

"Twitter is good for people, but it may make people less polite on the other hand it gives us a lot of information."

Suppose we try to analyze this statement according to point 1, the belief that the end state of having comfortable life is desirable, Also the reader further tries to prefer having comfortable life over being

polite. Hence we can say that human values will help in providing the context for categorizing, comparing, and evaluating argumentative statements which can help researchers in social science by getting informed with values through large-scale datasets, it can also help in selecting the arguments based on the target audience, assess arguments with scope and strength.

In this paper, we automatically identify human values in written arguments in this paper. For all taxonomic levels, the classification of human values is done using a variety of methods. For the purpose of conducting a comparative analysis, we implement, train, and assess many models. Results are given at the taxonomy level and demonstrate favorable results within and across cultures.

2 Background

2.1 Values in Social Science

A value system is the ranking of values based on cultural, societal, and personal considerations, according to [6]. Values are beliefs about desired end states or patterns of behavior. These concepts provide value to people rather than things, which makes methodical analysis easier[6]. These definitions are followed in the current work, which focuses on the values that personal arguments, most often implicitly, invoke. Fewer than hundreds of human values are thought to exist in total, according to [6], who also creates a useful survey of 36 values that makes a distinction between those that correspond to desired end states and those that promote good action. 48 value questions were generated by [2] specifically for cross-cultural study from the universal requirements of people and communities, such as following the law and being modest. Additionally, Schwartz (1994)[3] suggests that values are connected by their propensity to be compatible in their pursuit.

2.2 Values in argumentative research

The effectiveness of an argument depends on how highly the audience holds the values it invokes. Formal argumentation uses value systems to describe audience-specific preferences. Examples include defeasible logic programming [7], the value-based argumentation framework of BenchCapon, and value-based argumentation methods[8]. Aiming at the most crucial topics covered in a discussion, several researchers have looked at the problem of opinion summary in arguments[8][9][10]. Relatedly, the goal of key point analysis is to provide a limited number of succinct assertions, each of which represents a different element[11][12] . We contend that concentrating on the ”why” underlying an argument's logic, and assessing the values found in a group of arguments offers a fresh perspective on many areas of argumentation.

3 System Overview

We will be using the following models:

3.1 Bert

Google created Bidirectional Encoder Representations from Transformers (BERT), a transformer-based machine learning approach for pre-training natural language processing (NLP). At its core, BERT is a transformer language model with self-attention heads and a variable number of encoder layers. The architecture resembles Vaswani et alinitial .'s transformer implementation ”nearly exactly”. Language modeling (15 % of tokens were hidden, and BERT was trained to infer them from context) and next sentence prediction were the two tasks that BERT had been pre-trained on (BERT was trained to predict if a chosen next sentence was probable or not given the first sentence).

3.2 SVM

Support vector machines (SVMs, also known as support vector networks) are supervised learning models and learning algorithms that evaluate data for regression and classification purposes. SVM assigns training samples to spatial coordinates in order to maximize the distance between the two categories. Then, based on which side of the gap they fall, new samples are projected into that same area and predicted to belong to a category.

3.3 DistilBERT

It is a kind of smaller version of BERT, the main advantage of using DistilBERT is that it is a faster and lighter version of BERT. The important thing is even though it is lighter and faster it matches BERT's performance.

3.4 Roberta-base

is developed using a number of modifications to the original BERT architecture, including longer training times with larger data sets and batches, the removal of the objective to anticipate the next sentence, dynamic masking during pretraining, and the inclusion of a different tokenizer called byte-level BPE.

3.5 Bert tiny, small and medium

These 3 models are the various variations of BERT which are modified and used as an alternative for other Compact NLP pre-training models which are an alternative for huge and costlier NLP models which needs a huge volume of general-domain text. There are numerous model compression strategies on pre-trained language representation have been suggested due to the expense of this.

4 Data

In this section we will look at the dataset we used for studying human values behind arguments. Three crowd workers each annotated one of the 5270 and included arguments for each of the 54 values. we used different argument sources for different cultures. The dataset is divided into four sections in accordance with the ambition of a cross-cultural value taxonomy and using territories as a stand-in for cultures: China, India, Africa, and the USA[13]. Each argument has a premise, a conclusion, and a stance attribute that designates whether the premise is in favor or opposed to the conclusion. The dataset contains additional relevant arguments for the non-US regions because the existing argument datasets are nearly entirely from a Western background, thus reducing their size. This information is being used to train and compare classifiers across sources, not to represent the specific culture. The following table shows the no of unique conclusions and premises for each part of the dataset.

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

Crowdsourcing of Value Annotations: With the use of keyboard shortcuts and an obvious template-like structure, dataset consists of a unique three-part annotation interface that was optimized for speed and task skill acquisition. There are tree panels which are the main parts of interface. The first panel consists of arguments to be annotated in the scenario.

random person is arguing [in favor of/against] "[conclusion]" and stating: "[premise]."

In the second panel, the annotation task considers the statement as a yes/no question. for instance

adrie asks "are we good?", Is it's justification? "because it's good to have [value]"

The annotation progress is displayed in a third panel. Commentators could offer feedback on both values and arguments. For the dataset they used MACE[14], applying it value-wise as the author indicated for multi-label annotations, to combine the annotations into a single ground truth. Although the annotation assignment was challenging, the crowd-worker annotators managed to attain an average value-wise agreement of 0.49.

5 Experimental setup

5.1 Models Used

we used various models as indicated in the above section, You can find the implementation details

https://github.com/rohittaware1997/NLPTeam24

following are the various parameters that we used with different models that we used:

SVM we used linear kernel sci-kit-learn support vector machine which is trained label-wise with C = 14, 18, 20 and 22, and we used 5000 and 8000 iterations.

BERT we used bert-base-uncased with batch size of 12 and learning rate varing between $1^3 - 2^2$ and 15 epoches.

BERT-tiny we used BERT-tiny with a batch size of 12 and a learning rate varying between $1^3 - 2^2$ and 15 epochs.

BERT-small we used bert-base-uncased with batch size of 12 and learning rate varing between $1^3 - 2^2$ and 15 epoches.

RoBERTa we used batch size 12 and learning rate varing between $1^3 - 2^2$ and 15 epoches.

DistilBERT we used batch size 12 and learning rate varing between $1^3 - 2^2$ and 15 epoches.

5.2 Evaluation

We will use various parameters like F1-Score, precision, recall and mean across all labels for evaluation. We will also be giving the accuracy for completeness as we are using skewed label distribution which is making this accuracy less ideal. The matrices that we are using for evaluation, we will be using means to give each value the same weight. BERT is very effective as the recall is always near 1.

6 Results

There are total of 3 tables, **table 1** indicates F1 scores over each testset over all labels, **table 2** represents accuracy over all labels that we run over the various **BERT model**, **table 3** represents F1 Scores over each testset over all labels for **SVM** for various iterations. These results are for batch size of 12 with epochs 15. Please refer to results in the tables below. You can find the predictions under predictions folder in below repo.

https://github.com/rohittaware1997/NLPTeam24

First, we report results for the main part of the dataset (US) as an experiment with matched training and test sets. The approach is trained on arguments out of 60 distinct conclusions (2340 arguments, 83%), validated on 4 (3, 6%), and tested on 7 (400, 11 %). Conclusions were chosen to include approximately the proportions of arguments indicated in the various sentences. Unfortunately, this process resulted in different distributions of values in different sets. However, we wanted to test whether the classifier generalizes to unseen inferences, so we thought splitting the inferences was more important for our experiment. Only one very rare value, be neat and tidy (0.2% of arguments in the USA part), does not occur in the test set. We thus exclude this value from evaluation.

The relatively poor performance at higher levels is somewhat surprising, as it indicates that the categories at these higher levels are more difficult to separate using modern language-based approaches. is. Perhaps hierarchical classification approaches can address this relatively poor performance by using signals simultaneously at each level of the hierarchy. The F1 score of 0.25 at level 1 is encouraging for the 's rather unconventional approach, but the needs a lot more work.[1] A recall of 0.19 might be acceptable for a completeness-independent application, but precision of 0.40 is clearly too low for a real application.

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 in the table. However, this deficiency equally impacts the BERT, allowing for a comparison with the prior arrangement.

Even while more information and research seem to be required to reach this conclusion, these findings provide the first evidence that adopting a cross-cultural value taxonomy could produce reliable ways for determining the values behind arguments.

7 Conclusion

It is difficult to computationally identify the human values that underlie arguments. Our study makes a contribution of empirical analyses that compare various cultures and span many value granularity levels. We used various models like BERT, BERT-tiny, BERT-small, SVM, 1-baseline, distillery-base-uncased, and Roberta-base. This paper provides empirical studies that compare distinct cultures and span multiple value granularity levels using a dataset with a multilevel taxonomy of 54 values and 5270 arguments from four sources.

Even if more data and research appear to be needed to support this conclusion, the findings of this paper show that using a cross-cultural value taxonomy could result in dependable methods for identifying the values supporting claims. Based on this study, doing analyses that fully leverage label relationships is the logical next step. We can

[1pt] 1*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 pt] BERT	0.23	0.24	0.28	0.41	0.40	0.44	0.37	0.63	0.71	0.74	0.74	0.85	0.91	0.84	0.95	0.95	0.95	0.95	0.93	0.99
BERT-tiny	–	–	–	–	0.00	0.02	–	0.0	0.68	0.64	0.63	0.69	0.93	0.82	0.88	0.84	0.97	0.93	0.92	0.82
BERT-small	0.15	0.18	0.15	0.18	0.30	0.30	0.27	0.29	0.73	0.64	0.61	0.68	0.94	0.83	0.89	0.84	0.98	0.94	0.93	0.92
RoBERTa	0.24	0.23	0.24	0.27	0.34	0.38	0.34	0.35	0.73	0.68	0.67	0.71	0.94	0.82	0.91	0.84	0.97	0.94	0.94	0.92
DistilBERT [1.5pt]	0.20	0.20	0.20	0.27	0.30	0.33	0.29	0.34	0.71	0.68	0.61	0.69	0.93	0.83	0.88	0.83	0.96	0.93	0.92	0.91

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

[1pt] 1*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 pt] BERT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	10
BERT-tiny	0.93	0.93	0.93	0.95	0.83	0.86	0.88	0.90	0.71	0.74	0.58	0.71	0.86	0.74	0.79	0.75	0.93	0.86	0.85	0.82
BERT-small	0.93	0.92	0.94	0.95	0.85	0.86	0.88	0.89	0.74	0.72	0.54	0.72	0.86	0.75	0.78	0.74	0.93	0.86	0.85	0.82
RoBERTa	0.93	0.91	0.94	0.95	0.86	0.85	0.76	090	0.74	0.73	0.58	0.72	0.87	0.75	0.81	0.75	0.92	0.86	0.85	0.82
DistilBERT [1.5pt]	0.93	0.91	0.93	0.95	0.85	0.86	0.87	0.91	0.73	0.76	0.61	0.72	0.87	0.76	0.79	0.74	0.92	0.86	0.85	0.8

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

use various methods here like hierarchical classification methods[15]. We can also use multi-label categorization earning rules to reveal information about value linkages.

The power of an argument is significantly influenced by values, and substantial web data mining could improve all stages of argument formation, evaluation, and classification[16]. For example, it can be helpful to contrast the merits of opposing and supporting arguments. Declaring the values supporting arguments in unambiguous language could help prevent misunderstandings between people and automated argumentation systems[17]. Similarly to that, a beginning step toward resolving disagreements that appear to have extremely deep roots could be an "objective" analysis of the shared ideas that underlie differences between political parties.

Various researchers who are doing active research in the area of social science are very interested in examining values in huge text corpora. Which are the values communicated online? we can keep the track of references to the values over time by using the data in specified time archived over the internet. This research can further be extended for research into how the general public experiences human values in a such fast pace and digital era.

8 References

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

[2]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)

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

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

[5]Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, Benno Stein, 2022. Identifying the Human Values behind Arguments, Association for Computational Linguistics

[6] Milton Rokeach. 1973. The nature of

[1.5pt] 1*Country	C=14					C=18					C=20					C=22				
	1	**2**	**3**	**4a**	**4b**	**1**	**2**	**3**	**4a**	**4b**	**1**	**2**	**3**	**4a**	**4b**	**1**	**2**	**3**	**4a**	**4b**
[1 pt] Africa	0.16	0.27	0.64	0.85	0.88	0.17	0.20	0.57	0.79	0.86	0.21	0.20	0.67	0.79	0.86	0.17	0.20	0.56	0.79	0.86
China	0.13	0.25	0.54	0.78	0.85	0.13	0.25	0.54	0.78	0.86	0.13	0.24	0.53	0.79	0.85	0.14	0.26	0.57	0.73	0.83
India	0.18	0.24	0.57	0.72	0.82	0.19	0.24	0.57	0.72	0.82	0.17	0.25	0.58	0.74	0.82	0.19	0.25	0.57	0.74	0.83
USA	0.17	0.27	0.64	0.85	0.88	0.16	0.27	0.64	0.85	0.87	0.18	0.26	0.64	0.85	0.88	0.16	0.27	0.64	0.85	0.88
[1.5pt]																				

Table 3:
Macro F1-scores on each test set over all labels by level using **SVM** with different C values and **Max-iterations as 8000**

human values. New York, Free Press.

[7] 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.

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

Charlie Egan, Advaith Siddharthan, and Adam Z. Wyner. 2016. Summarising the points made in online political debates. In Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany. The Association for Computer Linguistics. [10]Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

[11]Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In Proceedings of the 58th Annual

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

[13]Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In 34th AAAI Conference on Artificial Intelligence (AAAI 2020), pages 7805–7813. AAAI Press. [14]Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), pages 1120–1130. Association for Computational Linguistics.

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

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

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