Rudolf Christoph Eucken at SemEval-2023 Task 4: An Ensemble Approach for Identifying Human Values from Arguments

Sougata Saha, Rohini Srihari
State University of New York at Buffalo
Department of Computer Science and Engineering
{sougatas, rohini}@buffalo.edu

Abstract

The subtle human values we acquire through life experiences govern our thoughts and gets reflected in our speech. It plays an integral part in capturing the essence of our individuality and making it imperative to identify such values in computational systems that mimic human actions. Computational argumentation is a field that deals with the argumentation capabilities of humans and can benefit from identifying such values. Motivated by that, we present an ensemble approach for detecting human values from argument text. Our ensemble comprises three models: (i) An entailment-based model for determining the human values based on their descriptions, (ii) A Roberta-based classifier that predicts the set of human values from an argument. (iii) A Roberta-based classifier to predict a reduced set of human values from an argument. We experiment with different ways of combining the models and report our results. Furthermore, our best combination achieves an overall F1 score of 0.48 on the main test set.

1 Introduction

Human values (Searle, 2003) play an integral role in determining how and why people react and respond to events in a particular way (Schwartz, 1994). We acquire them throughout our lifetime, and they largely govern our actions and manifest through our arguments. They play a significant role in defining our individuality and show a glimpse of our inner self by revealing our belief system. Motivated by why people reason and respond in a certain way, we present computational models capable of automatically identifying such perceivable values from argument text.

Given a set of 20 human value categories, the task (Kiesel et al., 2023) comprises classifying the most likely value categories from textual arguments in English. Our approach includes an ensemble comprising three models: (i) An entailment-based model for determining the human values based on their descriptions, (ii) A Roberta-based classifier that predicts the set of human values from an argument. (iii) A Roberta-based classifier to predict a reduced set of human values from an argument. Compared to standard classification-based approaches like models (ii) and (iii), we observe superior results from the entailment-based model. We experiment with different ways of combining the models[1], attaining an overall F1 score of 0.48 on the main test set, which is an improvement over the baseline of 0.42.

2 Background

We used the standard training, validation and test splits of the shared-task dataset (Mirzakhmedova et al., 2023) in English, which is based on the Webis-ArgValues-22 dataset (Kiesel et al., 2022). The dataset comprises 9,324 arguments (5,393 train, 1,996 validation, and 1,935 test) from 6 diverse sources, covering religious texts, political discussions, free-text arguments, newspaper editorials and online democracy platforms. Each argument is manually annotated for 20 human value categories (L2) spanning 54 human values (L1). Given a textual argument and a human value category, the task involves classifying whether an argument draws on that category.

3 System Overview

Our implementation comprises an ensemble of three models, where a primary model performs textual entailment and two secondary models perform classical text classification.

3.1 Textual Entailment Model

Although the task constitutes identifying human value categories (L2) from argument text, the dataset also contains finer human values (L1) and

[1] We release our code here: https://github.com/sougata-ub/semeval_2023_rudolf_christoph_eucken

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 660–663
July 13-14, 2023 ©2023 Association for Computational Linguistics

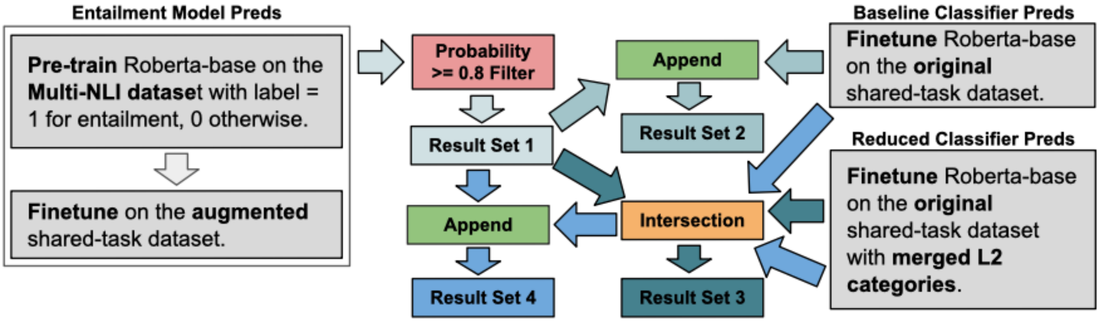
Entailment Model Preds
Pre-train Roberta-base on the Multi-NLI dataset with label = 1 for entailment, 0 otherwise.
Finetune on the augmented shared-task dataset.
Probability >= 0.8 Filter
Result Set 1
Append
Result Set 4
Append
Result Set 2
Intersection
Result Set 3
Baseline Classifier Preds
Finetune Roberta-base on the original shared-task dataset.
Reduced Classifier Preds
Finetune Roberta-base on the original shared-task dataset with merged L2 categories.

Figure 1: Ensemble Prediction Pipeline.

their textual descriptions. For example, the value category 'Self-direction: thought' represents the values 'Be creative', 'Be curious' and 'Have freedom of thought'. Each value is described by a set of sentences explaining what it means to possess such a human value. For example, 'promoting imagination' and 'being more creative' are two of the several descriptors for the value 'Be creative'. We implement an entailment-based model for identifying the value descriptors that an argument text entails. We transform the dataset conducive for textual entailment by creating argument and value descriptor pairs with its binary L1 label as the target to train our entailment-based model. We report the predictions aggregated at the L2 level.

Figure 2 illustrates our entailment dataset creation pipeline. Each example comprises an argument and value description pair and a binary label for entailment. The argument text is created by concatenating the premise, stance, and conclusion. The description text is created by prepending a value descriptor with its L1 label and the keyword 'by'. We yield N positive and N negative entailment examples for an argument, where N is the total number of value descriptors across all its positive L1 labels. Further, the N negative examples comprise equal proportions of 'easy' and 'difficult' samples, where a 'difficult' descriptor belongs to the same value category and different value, whereas an 'easy' descriptor is randomly sampled from a distinct value category altogether. The entailment dataset comprises 187,058 training and 65,900 validation pairs, which we use to train a Roberta-based (Liu et al., 2019) entailment model.

We follow a two-phased approach for training the entailment-based model. First, we pre-train Roberta-base on the MultiNLI (Williams et al., 2018) dataset. For our purpose, we binarize the labels by assigning 1 to the entailment class and 0 otherwise. We train the model for one epoch using mini-batches of 32 examples and a learning rate of 1e-5. The model is optimized using AdamW (Loshchilov and Hutter, 2017) and achieves a macro F1 score of 0.91 on the combined standard validation splits (validation matched and validation mismatched). Finally, we fine-tune the resultant model on our curated entailment dataset to engender the final model. We use the same settings as pre-training and achieve a macro F1 score of 0.79 on our validation split.

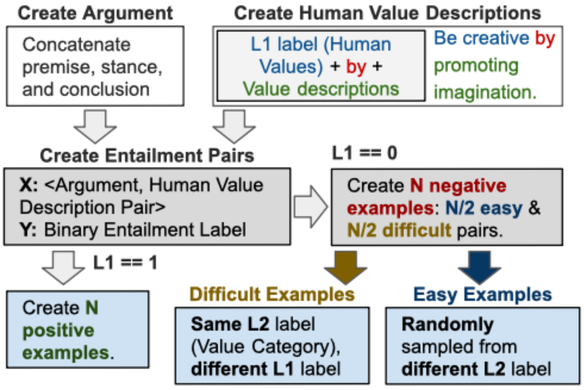
Create Argument
Concatenate premise, stance, and conclusion
Create Human Value Descriptions
L1 label (Human Values) + by + Value descriptions
Be creative by promoting imagination.
Create Entailment Pairs
X: <Argument, Human Value Description Pair>
Y: Binary Entailment Label
L1 == 0
Create N negative examples: N/2 easy & N/2 difficult pairs.
L1 == 1
Create N positive examples.
Difficult Examples
Same L2 label (Value Category), different L1 label
Easy Examples
Randomly sampled from different L2 label

Figure 2: Textual Entailment Dataset Creation Pipeline.

3.2 Textual Classification Model

We create argument texts as previously discussed, and train two variants of Roberta multi-label classifiers for directly predicting the perceivable value categories: (i) **Baseline Classifier:** We use the original 20 value categories as the target variable. (ii) **Reduced Classifier:** We split value category labels using ':' as the delimiter and combine the value categories starting with the same prefix, yielding a set of 12 classes. For example, the classes 'Self-direction: thought' and 'Self-direction: action' are combined into one class as 'Self-direction'. Leveraging Roberta-base as the base model, both classifiers uses a linear layer for predicting the target

Test set / Approach	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
Main																					
Best per category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
Result Set 1	.47	.40	.60	.20	.23	.60	.41	.49	.25	.70	.57	.61	.50	.46	.14	.49	.24	.71	.80	.20	.37
Result Set 2	.48	.39	.62	.20	.23	.61	.43	.48	.25	.70	.57	.62	.54	.46	.13	.49	.24	.71	.80	.26	.43
Result Set 3	.33	.38	.36	.20	.07	.26	.07	.37	.15	.51	.11	.09	.19	.09	.13	.49	.23	.71	.80	.20	.36
Result Set 4	.48	.40	.60	.20	.23	.60	.41	.48	.25	.70	.57	.61	.51	.46	.13	.49	.24	.71	.80	.26	.42

Table 1: Achieved F_1-score of team rudolf-christoph-eucken per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. 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.

classes from the pooler output. We train both models for 30 epochs with mini-batches of 64 examples and a learning rate of 2e-5. We use AdamW as the optimizer and stop training if the validation loss doesn't improve for 4 epochs.

4 Experimental Setup

Using each of the three models independently, we perform predictions on the standard shared-task validation and test sets. Illustrated in Figure 1, we combine the independent results using the following four schemes to generate the final set of predictions: (i) **Result Set 1:** Using the entailment model, we predict on the test set. We compare each example with all the value descriptors to check for entailment and filter out predictions less than probability 0.8. We report the value categories of the final value descriptors. (ii) **Result Set 2:** To increase the system's recall, we append the test set predictions from the baseline classifier with the result set 1. (iii) **Result Set 3:** Contrary to result set 2, here we try increasing the system's precision by reporting only the intersecting labels between the result set 1 and the reduced classifier predictions. (iv) **Result Set 4:** For each test set example, we first compute the intersecting labels from the baseline and reduced classifier predictions and append with the result set 1. Across all runs, we use TIRA (Fröbe et al., 2023) as the evaluation platform. We compute and compare overall and value category-wise F1, precision, and recall for the standard validation and test sets.

5 Results

We report our test set results in Table 1. Although our system falls short of the best approach, it performs significantly better than BERT and 1-Baseline, achieving a leaderboard position of 17th out of 41 teams in TIRA. As depicted in Table 2, compared to only classification-based systems (Baseline and Reduced), the entailment-based approach independently yields superior validation results. Hence, we used the entailment-based model output as our base result set (Result Set 1) and leveraged the classification-based model outputs to improve the entailment-based results. As indicated in Table 1, we observe that appending the classification-based results enhances the performance on the test set (in Result Set 2 and 4), whereas filtering out the entailment-based results based on the classification results (in Result Set 3) hampers performance.

Model	F1	Precision	Recall
Textual Entailment	0.49	0.44	0.56
Baseline Classifier	0.39	0.56	0.30
Reduced Classifier	0.26	0.17	0.53

Table 2: Overall model comparison on validation set.

We reason that since models like Roberta use large data samples for training, they learn to encode the semantics better: They can discern the meaning of an argument and associate it with the value descriptions. Furthermore, Roberta is pre-trained using the next-sentence prediction task and can associate relationships between sentence pairs. On the contrary, since classification-based approaches require learning domain-specific relationships, they are bottlenecked by the volume of the available training data. Hence, an ensemble approach such as Result Set 2 and 4 leverages the best of both scenarios and yields the best results.

6 Conclusion

Computational models for automatic detection of human values from arguments is a nascent yet crucial research direction for enabling purposeful argumentation systems. Here we present an ensemble comprising entailment and classification-based models for detecting human values from argument text. We experiment with different ways of combining the model outputs and attain an overall F1 score of 0.48 on the main test set, which significantly improves the baseline.

References

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 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. 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.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.