Human Value Detection

April 26, 2024

Abstract

Human Value Detection which is Task 4 for SemEval 2023 aims to predict the human values from a set of 20 value categories in a given argument. The nuanced, implicit and often subjective nature of value categories in textual arguments makes this a difficult task. The slightly small size of the dataset and lack of label-awareness adds an added challenge. To address these challenges we propose a model which draws on Textual Entailment and Contrastive Learning. The model acheives an F1 score of 0.5681 which beats the best submission in SemEval 23. The code is available on GitHub

1 Introduction

Argumentation is used almost everywhere in daily life from political canvassing to advertising campaigns. The kinds of arguments given by individuals is reflective of their own value systems.Moreover, the way a listener perceives and interprets an argument depends a lot on the value system of the listener. As a result two different listeners may interpret the same argument completely differently.

Consider the following arguments:
1. Capital punishment has been proven ineffective in reducing crime rates.
2. There are too many historical cases where innocent people have been put to death.
3. Capital punishment should be removed from society as it is a backwards and medieval solution to crime that has no place in a modern society.

Each argument presented here supports the abolition of capital punishment but from different value-driven perspectives:

The first argument advocates for abolition based on its ineffectiveness at reducing crime rates, relying on the value of Universalism: objectivity. This emphasizes rational, scientific thinking and an unbiased approach, illustrating the reliance on objectivity. The second argument combines objectivity with values of care, concern, and justice, highlighted through the use of "innocent person," invoking a compassionate and equitable standpoint. The third argument portrays capital punishment as outdated, advocating for more humane methods consistent with modern values. It draws on Face (protecting public dignity), Benevolence: Caring, Security: Societal, and Universalism: Concern, suggesting a multifaceted value approach emphasizing human dignity, societal safety, and progressive ideals.

While diverse in their value foundations, these arguments converge on a common goal: abolishing capital punishment. This complexity underscores the necessity for automated systems capable of dissecting such nuanced arguments to aid in understanding underlying values without explicit mentions.

This nuanced approach to argumentation enhances chatbot technologies by enabling more personalized and contextually appropriate responses based on user values. It also informs political campaigns and marketing strategies, allowing for tailored messages that deeply resonate with different population segments, increasing persuasion and engagement. Additionally, it has significant educational implications, helping educators develop students' critical thinking skills by teaching them to recognize underlying values and biases in arguments, fostering more open and constructive discussions.

We work on **Task 4 of SemEval 2023: Given a textual argument and a value category classify whether or not the argument draws on this value.** We also generate explanations as to why certain arguments draw strongly upon certain values using short descriptions for each of the values.

2 Related Work

2.1 Loss Function

Ma et al., 2023 [1] explore various loss functions and highlight Class Balanced-Negative Tolerant Regularization (CB-NTR) as the most effective. This approach uses a re-weighting strategy based on the effective number of samples per class to mitigate bias towards overrepresented classes. Additionally, NTR addresses the issue of negative label over-suppression in our task by adjusting their impact in the loss function, ensuring a balanced influence on the model's learning.

2.2 Contrastive Learning

Contrastive learning enhances PLM representations by promoting similarity among related examples and diversity among unrelated ones. As detailed in [2], this approach integrates a contrastive regularizer into the BCE loss, forming a combined loss function $L = L_{\text{BCE}} + C_{L_{\text{reg}}} \times \lambda$ where λ set to 0.1. The method generates two representations from one input, using contrastive loss to minimize distances between similar pairs and maximize those between dissimilar pairs, thereby improving PLM performance by addressing typical isotropy issues.

2.3 Textual Entailment

Kierkegaard et al., 2023 [3] proposed transforming the task from multi-label to binary classification by reframing it as a Natural Language Inference (NLI) task. In this approach, the argument is used as the premise and descriptions of human values serve as hypotheses. This method not only allows the use of pre-trained entailment models but also increases the number of instances available for training. It also increases the models ability to assign labels effectively by making the model label-aware.

2.4 Data Augmentation

Monazzah and Eetemadi, 2023 [4] present a methodology to improve human value detection in text using data augmentation. The technique leverages metadata describe labels within the dataset with the aim to enrich the training data with more information. They tested the effectiveness of their approach by fine-tuning pre-trained BERT and RoBERTa on the augmented dataset.

2.5 Ensemble Methods

Schroter et al., 2023 [5] proposed an ensemble of twelve transformer models that achieved the best results for a task at SemEval. These models were individually trained to minimize loss and were combined by averaging their predictions and setting a decision threshold based on a leave-out dataset. Although this ensemble outperformed other models in terms of F1 score, its high computational demands limited its deployment in low resource environments.

3 Dataset

We have used the **Touché23-ValueEval Dataset** [6] for Identifying Human Values behind Arguments which was publicly available. This dataset contains 9324 arguments collected from 6 diverse sources, covering religious texts, political discussions, free-text arguments, newspaper editorials and online democracy platforms.

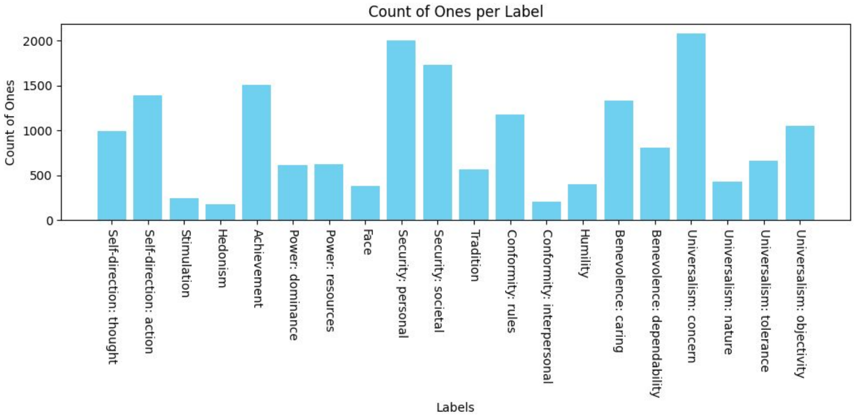


Figure 1: Label Distribution

Each argument in the dataset consists of one premise, one conclusion and a stance attribute indicating whether the premise is in favor(pro) or against(con) the conclusion.

The Dataset consists of 2 levels of annotations. The Level 1 annotations produced 54 labels which are more fine grained and which depicts the value whereas the level 2 annotations consists of 20 labels which are less fine grained and which depicts the value categories.

We have used the 20 labels(Level-2) for our task as was given in the SemEval'23 task.

This dataset have been divided into 3 sets as is customary in machine learning tasks. The 3 sets include the training set which consists of 61% of the samples, the validation set which consists of 21% of the samples and the testing set which consists of 18% of the total number of samples in the dataset.

4 Methodology

4.1 Textual Entailment

Drawing on the results of Kierkegaard et al. [3] we decide to model this as a textual entailment problem. For this the task is converted from a multi-label to a binary classification problem. We use a pre trained NLI model called pepa/roberta-base-snli. To solve for the class imbalance problem the instances from class 0 are dropped randomly. We also try using partial fine tuning of the model by fine tuning only the classifier head and the bias parameters. This helps reduce complexity and allows us to train for more epochs.

4.2 Conclusion Generation

We developed a novel dual-component model aimed at generating conclusions from texts given their premise, stance, and value labels. The first component is a classification model that predicts value labels from arguments, while the second, a generative model, generates conclusions based on the premise, stance, and predicted labels from the T5-small pretrained model. These models are trained sequentially: the classification model predicts the label, which is then used by the generative model to produce the conclusion. We hypothesize that combining the loss functions of both models can enhance classification accuracy. Due to computational constraints, the model was trained for only a few epochs.

4.3 Contrastive Learning

The main idea behind contrastive learning is to distinguish between pair of items. These pairs can be items that are similar or related in some way i.e. positive pairs or items that are dissimilar or unrelated i.e. negative pairs.

We have employed a contrastive learning approach leveraging the RoBERTa tranformer model to classify argumentative texts. Our model integrates three instances of the RoBERTa model to independently process the premise, conclusion, and stance. The outputs from these models (specifically, the [CLS] token representations) are concatenated along with the numerical "value" feature. This concatenated vector is then fed into a fully connected layer, which outputs a single value passed through a sigmoid activation function to obtain a probability, indicating the likelihood of the target class.

4.4 Ensembling

We observed a nuanced trend in the value categories; some were more specific variations of a broader category, such as 'Universalism: tolerance' and 'Universalism: objectivity' under 'Universalism.' To leverage this, we trained two 'roberta-base' models—one with all labels and another with a condensed label set. We then used an ensemble of these models, averaging their predictions for final inference accuracy.

4.5 Explainability

Explainability of model predictions is extremely important to understand any hidden biases and places where the model could improve. We have a t5 model which takes as input the entire argument along with all the true value categories and their

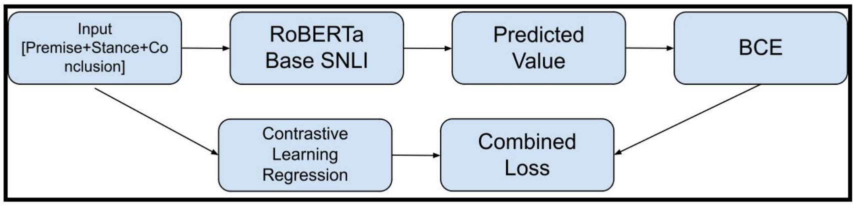

Input [Premise+Stance+Co nclusion]
RoBERTa Base SNLI
Predicted Value
BCE
Contrastive Learning Regression
Combined Loss

Figure 2: Model Diagram

descriptions. This model is run for 5 epochs. We pass the arguments and predicted value categories from the validation set.

4.6 Entailment and Contrastive Learning

We propose a novel model that integrates textual entailment with contrastive and standard classification losses. By leveraging contrastive learning, the model acquires robust and discriminative embeddings, enabling it to accurately identify similar input pairs that require similar predictions and distinguish those that are fundamentally different.

By framing the challenge as a textual entailment task, we harness pre-trained weights from the roberta-base-snli model, enhancing our training dataset significantly. To address class imbalance, we adopt a sampling strategy from the majority class(0), similar to approaches used in textual entailment models. The input premise to the model is the concatenated premise,stance and the conclusion, while the hypothesis is the value description.

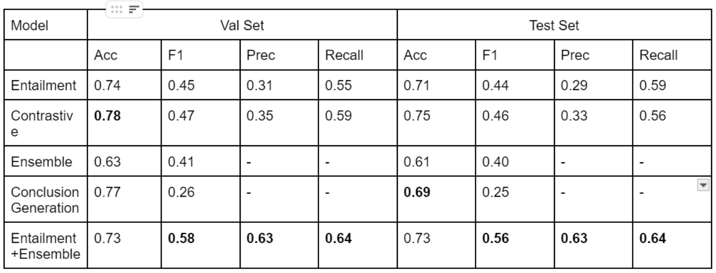
We anticipate that this model will exhibit strong performance on unseen data and effectively generalize to value categories with limited instances, due to its sophisticated understanding of input relationships fostered by the hybrid loss approach.

5 Experimental Setup

We use the Roberta-base-snli pre-trained model. The optimiser used was AdamW with a learning rate of 2e-5. Contrastive Learning Loss function was used. The model was trained for 3 epochs.

6 Results and Findings

We have compiled the results in Table 1. We have evaluated our code on four datasets (provided by the organiser): 1) Validation set obtained from the training



Model	Val Set				Test Set			
	Acc	F1	Prec	Recall	Acc	F1	Prec	Recall
Entailment	0.74	0.45	0.31	0.55	0.71	0.44	0.29	0.59
Contrastive	**0.78**	0.47	0.35	0.59	0.75	0.46	0.33	0.56
Ensemble	0.63	0.41	-	-	0.61	0.40	-	-
Conclusion Generation	0.77	0.26	-	-	**0.69**	0.25	-	-
Entailment +Ensemble	0.73	**0.58**	**0.63**	**0.64**	0.73	**0.56**	**0.63**	**0.64**

Figure 3: Results

set itself 2) Testing set 3) Nahj-al-Balagha, an Islamic religious text 4) New York Times
The following results show that our model with a test set F1 score of **0.5681 beats the previous best F1 of 0.561** set by Adam Smith. We test the robustness of the model using the Nahj-al-Balagha dataset and find the model gets an F1 of 0.46 more than the best **F1 of 0.36(Adam Smith)**. The **accuracy, precision and recall on the same dataset were 0.64,0.54 and 0.62** respectively. This shows the robustness of our model to unseen data.

We also analyse the F1 score for each value category and find the model fares poorly for Conformity-rules(0.39) and benevolence-caring(0.40). These labels were in a minority in the training set which could suggest the poor performance.

The explanation model mostly produced highly fluent explanations but was only very average in the adequacy. At times, the model's explanations, while displaying a high degree of fluency, showed a divergence in the adequacy of content. This occasional discrepancy between the value labels predicted by the classification model and the subsequent explanations may account for the observed variation in adequacy. The inter-annotator agreement on the quality of adequacy measured using Cohen's Kappa was calculated to be 0.7736 among two annotators. The fluency was by and large mostly good and consistent across different samples. The Cohen's Kappa for fluency was not calculated due to paucity of time.

7 Discussion, Analysis, Observations

The model also generalises better to unseen samples as we can see by our much improved results on the Nahj-al-Balagha dataset which has a completely different distribution to the main testing set.

We also expected the model to perform better for classes with fewer instances since the contrastive loss function can learn relationships in data beyond those between the input and labels. While we did observe some improvements compared to the other models as well as those proposed during the SemEval task the performance for these labels remained poor relative to other labels.
Explainability

8 Conclusion

In conclusion, our project aimed to predict value categories within arguments by using a binary classification approach based on the pre-trained roberta-base-snli model. This approach, enhanced with contrastive loss, improved our model's ability to analyze complex argument structures, achieving notable F1 scores that surpassed previous benchmarks.

We also developed a secondary model to generate explanations for each prediction, which, despite showing promise, requires further refinement, particularly in explanation adequacy. Moving forward, we plan to extend the capabilities of this model. We will focus on optimizing the computational efficiency to overcome current limitations and enhance the fluency and adequacy of the generated explanations.

Additionally, we will explore data augmentation techniques to better address class imbalances, which often challenge the training of robust models. Implementing reinforcement learning strategies could also play a pivotal role, potentially enhancing the adaptability and accuracy of our model under varying conditions.

Our efforts will also include developing methods to generate local explanations, providing users with more granular insights into the decision-making process of the model. These advancements aim to not only boost the model's performance but also its transparency and utility in real-world applications, ensuring it delivers comprehensible and contextually relevant insights.

References

[1] L. Ma, Z. Sun, J. Jiang, and X. Li, "Pai at semeval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 256–261, 2023.

[2] M. M. Oskuee, M. Rahgouy, H. B. Giglou, and C. D. Seals, "Tm scanlon at semeval-2023 task 4: Leveraging pretrained language models for human

value argument mining with contrastive learning," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 603–608, 2023.

[3] I. T. Cepeda, A. Pauli, and I. Assent, "Sren kierkegaard at semeval-2023 task 4: Label-aware text classification using natural language inference," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 1871–1877, 2023.

[4] E. M. Monazzah and S. Eetemadi, "Prodicus at semeval-2023 task 4: Enhancing human value detection with data augmentation and fine-tuned language models," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2033–2038, 2023.

[5] D. Schroter, D. Dementieva, and G. Groh, "Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models," *arXiv preprint arXiv:2305.08625*, 2023.

[6] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, B. Valentin, D. Dastgheib, O. Ghahroodi, M. A. Sadraei, *et al.*, "The touch\'e23-valueeval dataset for identifying human values behind arguments," *arXiv preprint arXiv:2301.13771*, 2023.