Human Value Detection
Group 14 Report

Wu Renjie
21018220
rwuap@connect.ust.hk

Xu Shengyu
20979790
sxubv@connect.ust.hk

Cui Xue
20973253
xcuiah@connect.ust.hk

Xie Linni
21003158
lxieaz@connect.ust.hk

Abstract

Human values are fundamental in forming individual viewpoints and content interpretations. Comprehending these principles is essential for numerous fields, such as natural language processing and social science research. With a particular emphasis on Task 4 of SemEval-2023, this article provides a method that makes use of pre-trained models and adds mid-level semantic categories, eventually achieving an F1 score of 0.49 on the main test dataset. Our model shows competitive performance in most categories and even outperforms the top rank models in more challenging categories. This work demonstrates how well pre-trained models can be used to improve human value identification when paired with mid-level semantics.

1 Introduction

The importance of detecting human values in natural language processing cannot be overstated. These values heavily influence human decision-making, leading to varied interpretations and responses to the same evidence. Incorporating human values into computational linguistics offers a rich context for classifying, comparing, and evaluating argumentative assertions (Mirzakhmedova et al. (2023)). The task of identifying these values seeks to uncover fundamental values by analyzing arguments from diverse sources such as political discourse, religious literature, and newspaper editorials (Kiesel et al. (2023)). Accurate detection of human values in text has far-reaching implications for fields like sentiment analysis, content recommendation, dialogue systems, and information retrieval, enabling more nuanced and context-aware language processing.

The objective of SemEval-2023 Task 4 aims to categorize textual arguments based on their alignment with specific human values. This work could support social science research, enhance argument analysis, tailor arguments to specific audiences, and illuminate common and differing viewpoints on contentious issues. Furthermore, it deepens our understanding of the relationship between language and human values, thereby enriching our knowledge of persuasive communication and its underlying dynamics. In our approach, we adopted a pre-trained model inspired by the top-ranking methodology of Adam Smith. After getting in touch with their team, we can explore further based on their best-performing model, thereby leveraging their expertise and streamlining our efforts. We also integrated mid-level semantics represented by 54 level-one labels in our dataset. which was motivated by our assumptions that intermediate-level semantics offer more specific and interpretable representations. This idea is similar to using additional features to enhance image classification. Our goal was to improve the detection of human values in textual arguments.

2 Background

The dataset provided by (Mirzakhmedova et al. (2023)) for detecting human values comprises 9324 arguments sourced from six distinct sources, including political debates, newspaper editorials, religious texts, and online democracy platforms. All arguments are in English and consist of three components: a stance attribute indicating whether the premise agrees ("in favor of") or disagrees ("against") with the conclusion, and two brief texts (the conclusion and the premise). The dataset is divided into two parts: the main dataset with 8865 arguments (95%) and a sup-

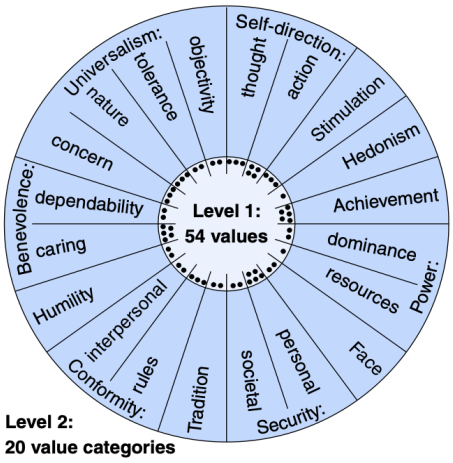

Figure 1: Human Values

plementary dataset with an additional 459 arguments. Each argument was annotated by three crowd workers for a total of 54 values (Figure 1), which were subsequently mapped to 20 value categories.

The Task Organizer, (Kiesel et al. (2023)), developed two baseline models: a BERT-based system and a 1-baseline. Our proposed system is fundamentally built on the DeBERTa model (He et al. (2020)), an enhanced version of the BERT model. The DeBERTa model incorporates disentangled attention and an improved mask decoder, enabling it to outperform RoBERTa-Large in various NLP tasks and achieve significant improvements in pretraining efficiency.

3 Related works

The majority of teams participating in SemEval 2023 Task 4 developed models to categorize whether a given textual argument leans towards a specific human value category. Adam Smith's approach (Schroter et al. (2023)), which was the best in this task, involved training transformer-based models until either the minimum loss or maximum F1-score was achieved. Combining all the models and selecting a single global decision threshold that maximizes the F1 score resulted in the competition's top-performing system. In the end, their ensemble system achieved an F1-score of 0.56 for the main test dataset with the optimal threshold of 0.25.

On the other hand, some other teams, such as PAI (Ma et al. (2023)), proposed a general multi-label classification system. This approach acknowledges that textual content can simultaneously associate with multiple values, adding a layer of complexity to the classification task. They also try quite a lot of loss functions to find out the best one. While Adam Smith's approach excels in optimizing Transformer-based models with a focus on F1 scores, PAI's multi-label classification system stands out for its ability to recognize and accommodate the inherent complexity involved in associating textual content with multiple human values simultaneously.

Our approach aims to combine the strengths of both teams. We leverage model optimization techniques for classifying multiple labels, aiming for a more comprehensive and accurate detection of human values in text.

4 Methodology

This section aims to provide a detailed description of the top-performing proposal as a foundation for further research and development. We begin with a brief overview of the prediction process, followed by an in-depth discussion of each stage. The following steps are taken to make a prediction:

1. We take an input argument and combine the premise, stance, and conclusion, a design inspired by Adam Smith (Schroter et al. (2023)). Additionally, we merge the value categories and level-one labels with the argument to create a 74-dimension label.

2. The input is then fed into the neural networks. The output from each neural network (from the two distinct strategies) is a vector with 74 labels (20 categories and 54 level-one labels) that indicate whether the sample contains the associated label or not, expressed as "confidence" (values between 0 and 1).

3. With the average values for each of the 74 labels, we need to decide which labels to apply. For this, we use a threshold. The corresponding label is applied to the values in the vector that exceeds the threshold.

Premise	Whaling is a part of a great number of cultures
Conclusion	We should ban whaling
Stance	Against
Labels	['Tradition', 'Conformity:interpersonal']
Level-1 labels	['Be protecting the environment', 'Have a world of beauty']

Table 1: Example argument before data preprocessing

4.1 Data Preprocessing

Based on the dataset provided by the Task Organizer (Kiesel et al. (2023)), we modified the labeled dataset by merging the 54 labeled level-one values with the 20 identified categorical values and ended up with 74 labels in total. The premise, stance, and conclusion were then combined into a single text string. Finally, we merged the text strings with all 74 labels to create a new dataset.

We chose to include Level 1 labels in our approach because we believe these lower-level semantic categories, compared with the high-level semantic labels of human value categories can enhance the identification of human values. while the primary focus of the challenge is to classify arguments into higher-level human value categories, the inclusion of Level 1 labels provides a more detailed understanding of the textual content.

Motivated by the idea that mid-level semantics can support more efficient learning, similar to enriching image classification with specific features, we hypothesized that incorporating Level 1 labels could lead to a more sophisticated and context-aware model. These labels serve as additional inputs, providing the algorithm with more detailed and accurate features to identify human values in textual arguments.

4.2 Model Architecture

Our model architecture, depicted in Figure 2, combines advanced pre-trained transformer models with level 1 label ensembles to enhance human value identification. We utilize the powerful DeBERTa Transformer as our pre-trained model, leveraging its demonstrated context awareness abilities refined on numerous similar tasks. The 54 level 1 labels collectively contribute to the learning process by providing the model with more low-level semantic information, crucial for in-depth text property analysis. We designed the 54 level 1 labels not only as the target of the classification task but also use it as the extra inputs for higher-level semantic human value categories.

We applied two strategies based on the design idea. As shown in Figure 3, two distinct strategies are implemented, each offering a unique method for deriving the final class prediction output. Strategy 1 encompasses predictions for level 1 and intermediate categories, which is used to be the output prediction of Adam Smith's model. Activation functions are purposefully used to enrich their representations, which are ultimately concatenated to form hierarchical categories. In contrast, Strategy 2 focuses on keeping more features before the final prediction. After applying an activation function, the original output is concatenated with the activated level 1 categories. From the architecture image, we can easily find that Strategy 1 has one more compressed hidden layer than Strategy 2. We will delve deeper into the performance of these two strategies in the results section. Due to its dual-strategy architecture, the model can recognize and capture intricate patterns in text parameters, ensuring a sophisticated and contextually aware interpretation of human values.

4.3 Train Process

Ensuring model robustness during the training phase requires a delicate balance between strategies, hyperparameter optimization, and loss calculation methods.

$$Loss = \alpha Loss_{category} + (1-\alpha) Loss_{level1} - F1 \quad (1)$$

The model parameters are frozen and then unfrozen in two stages, allowing for flexibility in the learning process, and also helping fine-tune pre-trained models in different tasks. The loss function (1) plays a critical

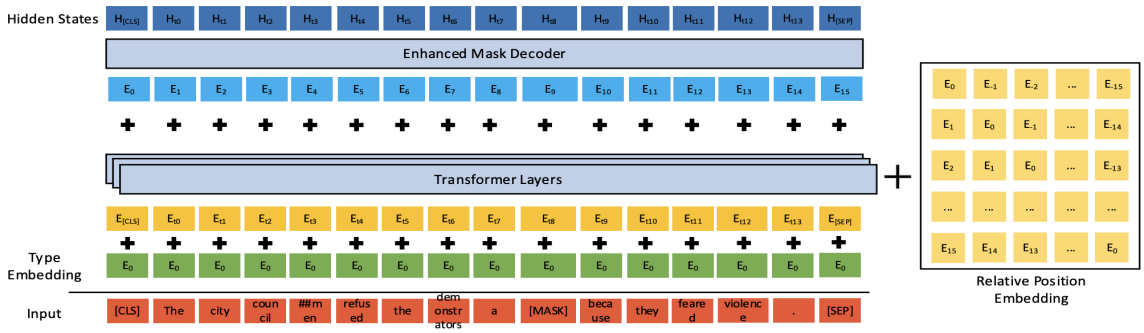

Figure 2: Model Structure of DeBERTa

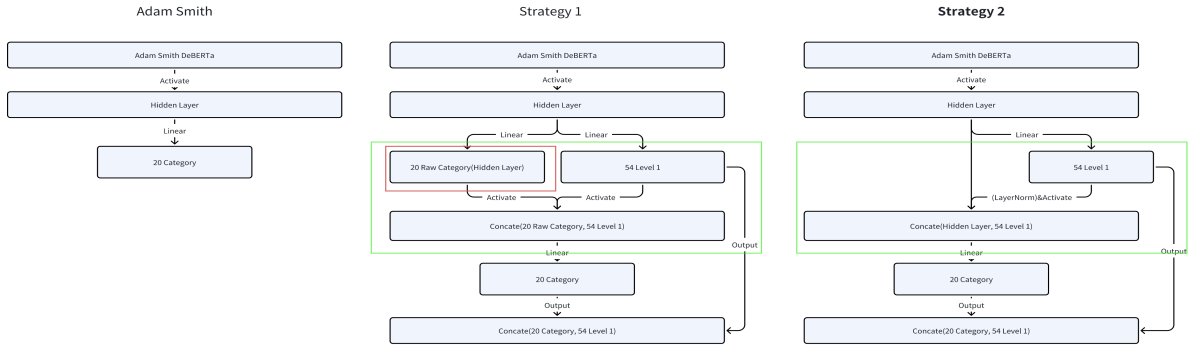

Figure 3: Model Structure of Our Strategies

role. We explore two additional approaches: total binary cross-entropy (BCE) (Domingos (2012)) and weighted BCE, with the latter incorporating policy weights based on the relative importance of different task components. Optimization approaches are tested by examining the convergence of the Adam and AdamW optimizers to the optimal solution. We also experiment with activation functions like ReLU and LeakyReLU (Nair and Hinton (2010)) to determine the best fit for our task. Dynamic threshold testing further refines the decision-making process after training, ensuring the model's adaptability to varying data complexities. This comprehensive training approach underscores the synergy between pretraining, low-level semantic enrichment, and optimal model configuration, aiming to develop a model that excels in nuanced human value recognition.

5 Experimental Setup

In our experimental design, we meticulously partition the dataset into distinct subsets to facilitate reliable training, validation, and testing of BERT models for human value identification. A validation set comprising 1896 samples enables us to test model performance and prevent overfitting, while a training set of 7998 data forms the basis for model learning and coverage. We reserve a specific test set to ensure a fair evaluation of the model's generalization capability. To address the issue of class imbalance, we employ a custom oversampling method to enhance the representation of minority instances, and data augmentation techniques are used to introduce diversity.

Our models' performance evaluation is based on a comprehensive set of metrics tailored for multi-label classification tasks. Binary cross-entropy loss serves as our primary loss function during training, quantifying the discrepancy between predicted and true labels. The F1 score, calculated based on each of the original 20 categories and macro-averaged across all categories, provides nuanced insights into precision and recall. Additionally, we measure the overall accuracy of the validation and test sets as an indicator of the proportion of correctly classified instances. Experiments with threshold optimization are conducted to explore further enhancements, ad-

justing categorization thresholds to strike a suitable balance between recall and precision. In the context of our experiments, additional metrics such as precision and recall for each category aid in providing a thorough evaluation of the model's effectiveness in human value detection.

All of our training is conducted on Colab A100 with 40GB GPU. The trend of validation loss and training loss is used to determine the direction of fine-tuning after each training process. The learning rate is set to 1e-5 for a warmup and set to 1e-6 for later processing. The default epoch is set to 30. The default threshold is set to 0.25 during the training. The number of train workers and validation workers are both set to 4.

6 Results and Analysis

6.1 Strategy Selection

As previously stated, at the start of the experiment, we trained each strategy several times. After a thorough examination of the performance indicators obtained from the validation dataset, we found that Strategy two overperforms Strategy 1 not only from F1 scores but also convergence speed. From Table 2, the F1 scores of strategy 2 consistently perform greatly better than strategy 1 for both the original 20 categories and the enlarged 54 level 1 labels. We think it may caused by the reason compressing of Adam Smith's DeBERTa to 20 features in the hidden layer leading to the information loss for final category prediction. So we finally chose strategy 2 as the architecture for the coming experiments.

6.2 Hyperparameter Tuning

Our experimental findings indicate that the choice of activation function has a significant impact on how well the model performs.

During the experiments of comparing different strategies, we found that strategy 2 beats strategy 1 in all aspects. Some of its concrete human value categories' F1 score is very low, even to be zero, as shown in Figure.4,5. We thought it may be caused by our default activation function ReLU, hence, We conducted tests comparing two popular activation functions for multi-label classification tasks: ReLU and LeakyReLU. The activation functions are tested on an enlarged set of 20 predefined categories and 54 first-level categories. The results of the experiments demonstrate that LeakyReLU does not relieve the problem of low F1 scores of some concrete human categories. It did outperform ReLU with a macro-average F1 score (0.61) better in the original 20 categories. This indicates that it is better able to represent the inherent complexity of these different human values, leading to more complicated and precise projections. Furthermore, LeakyReLU still performs well, with a macro-average F1 score of 0.35, considering the expanded 54 first-level categories (requiring broader predictions). This indicates how well it handled the complexity brought forth by the expanded collection of categories. LeakyReLU and class-specific fine-grained predictions are given priority in our model, which helps us select this activation function. This choice will direct the next stages of model optimization and hyperparameter modification, guaranteeing a well-balanced strategy appropriate for the difficulty of our classification problem.

We also thoroughly investigated the effects of decision thresholds, which range from 0.01 to 0.35, on model performance during our training. With careful consideration, we hope to uncover subtleties in the metrics of precision, recall, and F1 score, which will provide insights into how the model behaves at various thresholds to find its relationship with low F1 scores in concrete categories. In particular, a notable threshold of 0.03 is essential for striking a balance between recall and precision. Thresholds over 0.03 have high macro F1 scores in all 20 human value categories. These results underline the significance of our threshold decision, which was guided by a thorough investigation in line with the complexity of our dataset and the particular demands of the classification assignment. However, we still do not observe a strong relationship between thresholds with little part low F1 score categories.

An in-depth analysis of the data revealed significant differences in F1 scores and losses across categories. We have a total of 74 labels for training, and each of them takes the same importance. However, our final goal is to perform well in 20 human value categories, which may be decreased by the same weights during

Table 2: Metrics of Strategy 1 and Strategy 2 on 20 Categories

	Strategy 1			Strategy 2			
	precision	recall	f1-score	precision	recall	f1-score	support
Self-direction: thought	0.80	0.72	0.76	0.00	0.00	0.00	251
Self-direction: action	0.78	0.79	0.79	0.00	0.00	0.00	496
Stimulation	0.61	0.37	0.46	0.00	0.00	0.00	138
Hedonism	0.56	0.82	0.66	0.00	0.00	0.00	103
Achievement	0.86	0.78	0.82	0.00	0.00	0.00	575
Power: dominance	0.00	0.00	0.00	0.00	0.00	0.00	164
Power: resources	0.77	0.60	0.67	0.00	0.00	0.00	132
Face	0.00	0.00	0.00	0.00	0.00	0.00	130
Security: personal	0.85	0.85	0.85	0.88	0.79	0.83	759
Security: societal	0.86	0.72	0.79	0.00	0.00	0.00	488
Tradition	0.85	0.62	0.72	0.87	0.60	0.71	172
Conformity: rules	0.81	0.66	0.73	0.85	0.59	0.70	455
Conformity: interpersonal	0.60	0.43	0.50	0.00	0.00	0.00	60
Humility	0.43	0.28	0.34	0.00	0.00	0.00	127
Benevolence: caring	0.76	0.74	0.75	0.00	0.00	0.00	633
Benevolence: dependability	0.53	0.39	0.45	0.00	0.00	0.00	268
Universalism: concern	0.86	0.78	0.82	0.93	0.64	0.76	687
Universalism: nature	0.76	0.89	0.82	0.00	0.00	0.00	127
Universalism: tolerance	0.52	0.39	0.45	0.00	0.00	0.00	223
Universalism: objectivity	0.74	0.60	0.66	0.00	0.00	0.00	371
micro avg	0.78	0.67	0.72	0.89	0.22	0.36	6359
macro avg	0.65	0.57	0.60	0.18	0.13	0.15	6359
weighted avg	0.74	0.67	0.70	0.29	0.22	0.25	6359
samples avg	0.81	0.72	0.73	0.60	0.25	0.34	6359

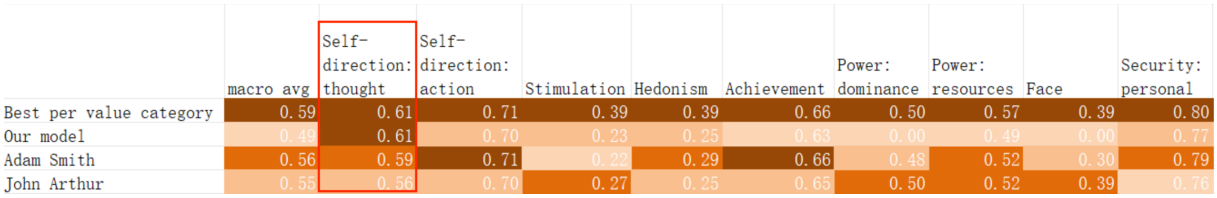

Figure 4: Comparison with top rank 2

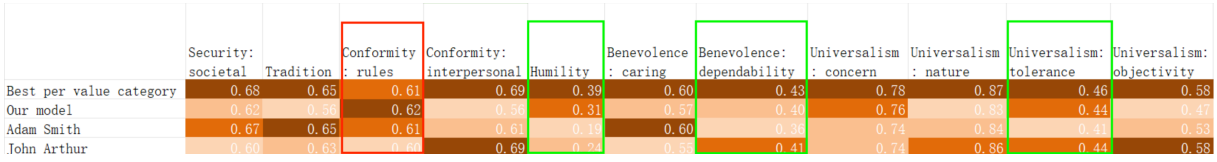

Figure 5: Comparison with top rank 2

training.

Consequently, our final choice is to use a weighted approach. This approach takes into account the different importance of different categories, ensuring that the model achieves balanced performance across different categories. The decision to use a weighted loss function was driven by considerations for our task requirements and is expected to help improve the overall performance of the model. We split the categories simply by their semantic levels, one is 20 human value categories, and another is 54 level 1 labels. We set different weights for them, hoping to emphasize the importance of 20 human value categories' loss. Also, learned from Adam Smith, that we

add training F1 score as part of our loss function, hoping that the increment of F1 score can also help reduce the final loss. We set different weights for 20 human value categories, using 0.8, 0.73, 0.68, 0.5. Finally, loss decreased and we got the best performance to 0.61 with all categories F1 scores more robust.

The evaluation of our final model underscores the effectiveness of our strategic choices, demonstrating strong performance across key categories. For more detailed information, please refer to the red columns indicated in Figure.4,5. A standout aspect is the model's proficiency in discerning subtle patterns within the Self-Direction: Thoughts category. Here, it matched the best performance with an accuracy of 0.61 and beat the top 2 models. Furthermore, our model excelled in Conformity: rules, achieving an impressive accuracy of 0.62 and even outperforming the best performance. Notably, our model has shown competitive accuracy in categories that are typically challenging for classification, such as Humility, Benevolence: Dependability, and Universalism: Tolerance. Our models all surpassed the top 2 models in those categories. These results underscore the robust performance and competitive edge of our model, often matching or surpassing the performance achieved by top competitors., thereby validating the effectiveness of our approach.

7 Conclusion

In conclusion, our journey into the realm of human values detection, specifically within the framework of SemEval-2023 Task 4, has been enlightening and fruitful. Our approach, inspired by Adam Smith's idea, leveraged pretrained models, a strategy that saves us a lot of computational resources, speeds up the training process, and even improves the performance of our model, especially when we don't have a lot of task-specific training data. The integration of mid-level semantics, represented by 54 level 1 labels, added a layer of specificity and interpretability to our detection process. After numerous attempts, we eventually achieved an impressive F1 score of 0.61, placing it among the top-tier models in this task. It demonstrated its efficacy across a wide range of human value categories, even those that are typically challenging for classification.

We focus on the methodical investigation of various models, loss functions, and parameters for feature work to improve the adaptability and efficiency of our human value identification system. We intend to test a variety of pre-trained models, going beyond BERT to include various transformer-based architectures that might be more adept at capturing the finely nuanced semantics found in textual arguments. Our feature study will also explore other loss functions, like focal loss and contrastive loss, to solve issues with imbalanced datasets and enhance recall and precision. Further research will focus on optimizing variables including learning rates, batch sizes, and activation functions to find the best setups for achieving the best results in a variety of human value categories. Our goal is to continuously improve our model through these endeavors, guaranteeing its flexibility to various linguistic expressions and cultural situations, and pushing the boundaries of human value recognition in natural language processing.

During the fine-tuning process, we embarked on a journey of exploration, experimenting with various elements such as loss functions, activation functions, and model structures to optimize performance. This journey not only led us to better performance but also deepened our understanding of state-of-the-art models and fine-tuning methodologies. However, we acknowledge that our model has its limitations. Notably, it achieved zero accuracy in the 'Power: dominance' and 'Face' categories. We attempted to address this issue through oversampling, but the results were not as promising as we had hoped.

Moving forward, it is crucial to investigate the root cause of this unusual performance. We believe that further research and experimentation will shed light on this issue and help us improve our model's performance in these categories. Our journey in the realm of fine-tuning is far from over, and we look forward to the insights and improvements that future work will bring.

References

Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

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 (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Long Ma, Zeye Sun, Jiawei Jiang, and Xuan Li. 2023. 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)*, pages 256–261, Toronto, Canada. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touche23-valueeval dataset for identifying human values behind arguments.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Daniel Schroter, Daryna Dementieva, and Georg Groh. 2023. Adam-smith at SemEval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 532–541, Toronto, Canada. Association for Computational Linguistics.