Tenzin-Gyatso at SemEval-2023 Task 4: Identifying Human Values behind Arguments using DeBERTa

Pavan Kandru, Bhavyajeet Singh, Ankita Maity, Aditya Hari and **Vasudeva Varma**
IIIT Hyderabad
{ siri.venkata, bhavyajeet.singh,ankita.maity,aditya.hari}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

Identifying human values behind arguments is a complex task which requires understanding of premise, stance and conclusion together. We propose a method that uses a pre-trained language model, DeBERTa, to tokenize and concatenate the text before feeding it into a fully connected neural network. We also show that leveraging the hierarchy in values improves the performance by .14 F1 score compared to only using level 2 values. Our code is made publicly available here.[1]

1 Introduction

Humans often come to different conclusions given the same premise. This variation can be attributed to their values. Identifying the values behind the arguments is helpful in understanding the argument itself. Downstream tasks like supporting or opposing argument generation can benefit from value identification. In this task(Kiesel et al., 2023), we aim to identify 20 value categories in a given premise, stance and conclusion pair. Data used is collected from four geographical cultures, manually annotated by (Mirzakhmedova et al., 2023).

This paper proposes a method for multi-label classification of the premise, stance, and conclusion pairs using Encoder only LMs as a background model. We use DeBERTa(He et al., 2021), a pretrained language model, that has shown remarkable success in various NLP tasks, including classification. The proposed method tokenizes the premise, stance and conclusion text using the pretrained tokenizer, and then concatenates them and feeds it into the LM, generates a representation of the combined text, and maps it to a set of values using a fully connected Neural Network. The model is trained on a Multi-margin loss function and evaluated on metrics such as accuracy, precision, recall, and F1 score. This approach can potentially result in a highly accurate and effective NLP model for identification of values in arguments.

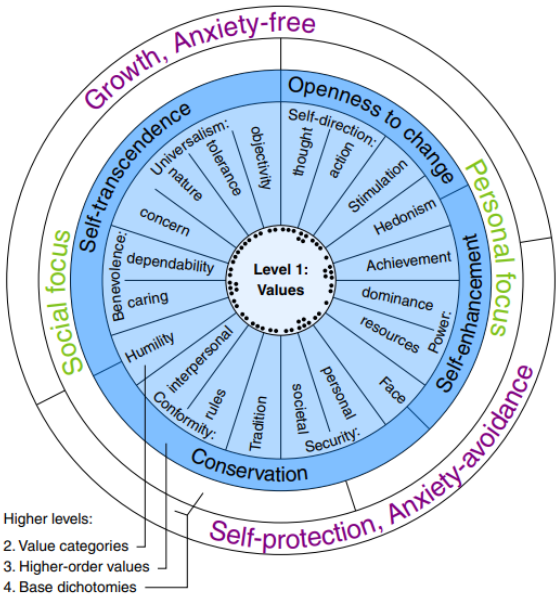

Figure 1: Values in the data organized higher level to lower level.

Our major finding from the task is that the hierarchy in the values helps with identifying them. We found that adding 5 high level values Self-direction, Power, Security, Conformity, Benevolence, Universalism to the existing 20 values improved performance substantially compared to just using the 20. These 5 values are extracted from labels of 20 values. For example, the value Self-direction: action belongs to the coarse Self-direction class. Our model outperformed the 1-Baseline and Random Baseline in (Fröbe et al., 2023) and performed similarly to the BERT Baseline.

2 Background

The task involves predicting what values would have lead humans to come to the given stance and conclusion when provided with a premise. There are a total of 20 such values that needs to be identi-

[1] https://github.com/pavankandru/values

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2062–2066

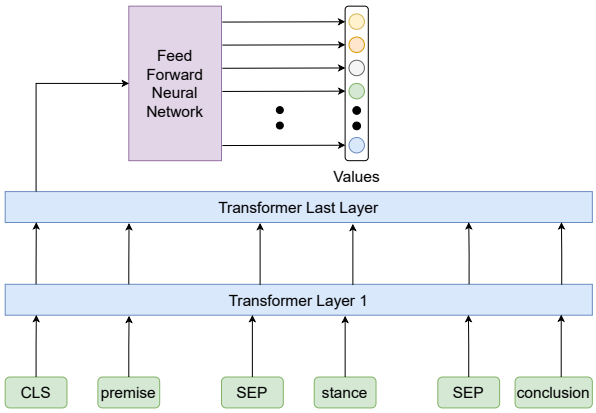

Figure 2: Values in the data organized higher level to lower level.

fied in an argument(premise, stance and conclusion pair). We are also provided with higher level values which are less nuanced than the required level to be predicted. Figure 1 shows the various levels of values present in the data, that are associated with arguments. We are interested in the level 2 values in this task.

Following is an example of a premise, stance and conclusion along with values associated with the argument. Stance is binary variable with 'in favor of' and 'against' as possible values.

premise: Surrogacy should be banned
stance: against
conclusion: Surrogacy should not be banned as it is the woman's right to choose if she wishes to do this for another couple and be compensated.
Level 1 values: Have freedom of action
Level 2 values: Self-direction: action
Level 3 values: Openness to Change
Level 4a values: Personal focus
Level 4b values: Growth, Anxiety-free

3 System Overview

We used an end to end neural approach to predict the values in an argument. Transformer Language Models have shown remarkable capabilities in NLP applications. We make use of Deberta model which is an encoder only model pretrained on data from Wikipedia (English Wikipedia), BookCorpus (Zhu et al., 2015), OPENWEBTEXT (Gokaslan and Cohen, 2019), and STORIES (a subset of CommonCrawl (Zellers et al., 2018)). Joint representations of premise, stance and conclusion are generated using Deberta. This is done by concatenating the premise, stance and conclusion with SEP token

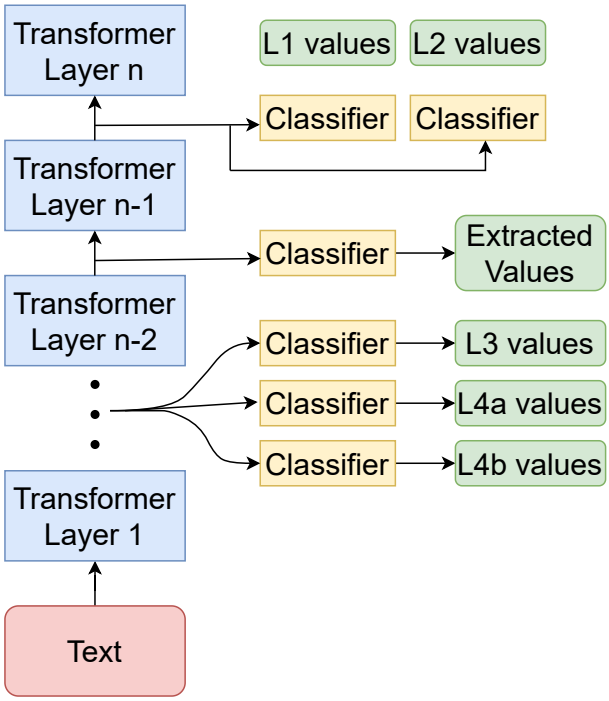

Figure 3: Using Internal Hidden states to feed classifiers to exploit the Hierarchy in Values

separating them. Concatenated text is then tokenized and fed into LM and last hidden state of CLS token is used as the representation of argument. This facilitates information exchange between premise,stance and conclusion while generating representation of argument. This representation then used as input to a Feed Forward Neural Network with dropout and SiLU activation function. FFNN maps the representation to 20 logits which corresponds to the values. We optimize Multi Margin Soft Loss to train the model end to end to perform Multi-label Classification. Following are some of the variations of this approach we tried.

3.1 Extra Classes

Values contain hierarchical information within them that can help us group some values into a cluster. We extracted 5 such clusters Self-direction, Power, Security, Conformity, Benevolence, Universalism and used these as extra classes while predicting the values.

3.2 Predicting More Levels of values

Values are hierarchical as seen in Figure 1. We hypothesise that Training models to predict other levels of values along with the ones we are interested in can help improve the performance on the level

2 value prediction. We used one classifier head for each level of values classification. Performance of model on values of each level are optimized using a corresponding loss function. Weighted average of lossses from different levels is minimized to train the model.

3.3 Using Hidden States of LM

Internal Hidden states of LM can be used to predict the auxillary losses mentioned in 3.2. (Szegedy et al., 2015) model proposed this to improve performance in deep Neural Networks. In a transformer with n blocks, we use nth hidden state to predict level 1 and level 2 values. n-1, n-2, n-3, n-4, n-5 th hidden states to predict the extra classes in 3.1, level3, level4a and level4b respectively.

4 Experimental Setup

We used the main dataset provided in the task website. It consists of a train split with 5393, a validation split with 1896 samples and a test split with 1576 samples. Level 2 values has 20 classes with high imbalance. The maximum class count to minimum class count ratio in train split is 12 and overall is approximately 10. This ratio for level 1 is 84, level 3 is 2, indicating that there is very high imbalance in levels that high number of values. Levels 4a and 4b has almost equal distribution of labels.

Level 3,4a and 4b values are absent for 2024 out of 5393(37%) arguments in training split. We fill the missing values with 0,1 values from an uniform random distribution.

Models mentioned in 3 are optimized using Adam optimizer with weight decay and learning rate 1e-4. DeBERTa model from huggingface(Wolf et al., 2019) is used and end to end model is trained with batches of size 16 without any gradient accumulation.

We use Precision, Recall and F1 score as metrics to evaluate the models as they are resistant to class imbalance problem when averaged using macro strategy. Scikit-learn's (Pedregosa et al., 2011) Implementation of macro measures are used.We use performance of models on val split to compare the models.

5 Results

Table 2 shows the results of our models on val split. Our Best approach on task official leader board is presented in Table 1. We can see that adding extra classes while prediction improved the overall F1 score by 14 points compared to not using them. Adding more levels of values reduced the performance slightly. This can be due to high percentages of missing labels for arguments. We have submitted the DeBERTa + Extra class model results and compared them against the test leader board. Our model performed similar to the BERT baseline approach proposed by the organisers.

6 Conclusion

In this paper we discuss the task to identify values in a given premise, stance, and conclusion pair, which can be useful in downstream tasks like generating arguments. We found that identifying the hierarchy in the values improves performance, and adding 5 high-level values to the existing 20 values significantly improved the model's accuracy compared to just using 20. Hierarchical methods did not perform as expected due to missing high level values. The proposed approach has the potential to be an effective NLP model for identifying values in arguments.

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.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. https://skylion007.github.io/OpenWebTextCorpus/.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

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.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib,

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
DeBERTa + Extra Classes	.41	.44	.62	.07	.12	.47	.33	.45	.10	.70	.57	.45	.42	.21	.12	.47	.21	.67	.73	.31	.37

Table 1: Achieved F_1-score of team tenzin-gyatso 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.

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
DeBERTa	.26	**.62**	.11	.12	.16	.07	.07	.07	.04	.00	.01	**.72**	**.62**	**.55**	**.46**	.00	.17	.54	.56	**.31**	.04
DeBERTa+Extra Classes	**.43**	.43	**.56**	**.25**	**.39**	**.59**	**.25**	**.37**	**.20**	**.70**	**.60**	.41	.49	.15	.08	**.52**	.21	**.63**	**.70**	.14	**.48**
DeBERTa+All Levels	.24	.23	.41	.00	.00	.47	.16	.13	.00	.57	.41	.17	.39	.00	.00	.50	.25	.53	.00	.21	0.33
Hierarchichal	.25	.31	.41	.00	.00	.44	.19	.20	.06	.54	.49	.18	.36	.00	.05	.40	**.26**	.56	.06	.22	.33

Table 2: Val split evaluation metrics for each of the 20 value categories.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.