Noam Chomsky at SemEval-2023 Task 4: Hierarchical Similarity-aware Model for Human Value Detection

Sumire Honda *
University of Potsdam
sumire.honda@uni-potsdam.de

Sebastian Wilharm
University of Potsdam
sebastian.wilharm@uni-potsdam.de

Abstract

This paper presents a hierarchical similarity-aware approach for the SemEval-2023 task 4 human value detection behind arguments using SBERT. The approach takes similarity score as an additional source of information between the input arguments and the lower level of labels in a human value hierarchical dataset. Our similarity-aware model improved the similarity-agnostic baseline model, especially showing a significant increase in the value categories with lowest scores by the baseline model.

1 Introduction

People often form different arguments given the same source of information. One of the factors behind the difference lies in their "different beliefs and priorities of what is generally worth striving for (e.g., personal achievements vs. humility) and how to do so." (Kiesel et al., 2022), which is referred to as human values (Searle, 2003). This paper describes our approach to human value detection through the submission to SemEval-2023 Task 4. (Kiesel et al., 2023).

Our main strategy is adding similarity scores between the argument inputs and the human values to predict which human value categories the argument draws on. The idea of the approach is to collaborate with the hierarchical relation of the input arguments towards the lower levels of more concrete human labels. We assume that such a hierarchical information will support the model to identify the attribute of the arguments within a context of the hierarchical human value dataset. The code is available at this link.[1]

In section 2, the background behind our proposed approach is described, and the approach method is explained in section 3 and section 4. Section 5 reports our results with F_1 score with discussions and error analysis.

2 Background

The goal of the task is to classify the human value categories with multi-labels, given textual arguments data with human value category labels. There are four different optional levels in the labels, and the level 2 labels serve as our prediction targets. For example, if the textual arguments would be *"The leaked personal information will be defrauded by fraud gangs to gain trust and carry out fraudulent activities"*, Our model should ideally predict the level 2 human value behind the arguments, *Self-direction: action*, *Security: societal*, and *Conformity: rules*.

The arguments of the dataset is collected from religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms, and they are written in or translated to English. The level 2 labels (20 labels) are annotated by crowd workers. Our team used the main dataset for training, validation and testing and a supplementary dataset Nahj al-Balagha for testing. Additionally, for our extended approach described in section 3, we utilized level 1 labels (54 labels of human values) and the example sentences for the level 1 labels to concatenate the hierarchical information to predict the level 2 target labels. The dataset details are described by Mirzakhmedova et al. (2023).

2.1 Morality and Human Values

A similar target to human values, namely moral sentiment detection, has been conducted with a computational approach; however, human value detection behind arguments has not been attempted before the study by Kiesel et al. (2022). For example, moral sentiment prediction on argumentative texts was studied by Kobbe et al. (2020). The idea of moral belief classified by moral founda-

*Equal contribution. Author order was determined by the alphabetical order

[1] https://github.com/su0315/SemEval23_Human_Values_Detection

tions (Haidt, 2012) is referred to as an important role in ideological debates. It cannot be resolved by simply comparing facts, and Feldman (2021) claimed it is strongly connected with human values. However, (Kiesel et al., 2022) indicated the difference between values and moral foundations by the vagueness of the foundations since moral foundations are categorized into 5 labels (care, fairness, loyalty, authority, and purity). In contrast to the moral foundations, human values consist of 54 labels which have hierarchical relation to the higher levels of categories.

2.2 Similarity for Augmentation

Misra et al. (2016) used similarity to label frequently paraphrased propositions or labels capturing the essence of one particular aspect of an argument called argument facets, such as morality. They extracted arguments on social media dialogues and ranked the arguments in terms of their similarity and demonstrated the potential of similarity to detect argument facets beating several baselines. Reimers and Gurevych (2019) proposed Sentence-BERT(SBERT), which is a modification of the pretrained BERT (Devlin et al., 2019) that can be compared via cosine-similarity given a pair of sentences. It uses siamese and triplet structures to derive semantically meaningful sentence embeddings so that the sentences are comparable with cosine-similarity. Although SBERT showed the potential of adapting to argument facet similarity corpus, the score decreased compared to other tasks. Therefore, Behrendt and Harmeling (2021) proposed ArgueBERT, where SBERT architecture is pretrained in several argumentation tasks, and the pretaining improved the performance on argument similarity.

2.3 Similarity information with human values to predict their higher-level categories

Our approach is utilising the hierarchical nature of the human value dataset (Mirzakhmedova et al., 2023) through similarity score concatenation to the baseline BERT model by Kiesel et al. (2022). The concatenated similarity score shows the semantic similarity between the premises of the arguments and the human values, which are lower level than our prediction target, level 2 human value categories.

We hypothesize that the similarity score will improve the level 2 human value categories detection from the premises, since the premises (ex. *The leaked personal information will be defrauded by fraud gangs to gain trust and carry out fraudulent activities*) are more concrete statements than the relatively abstract level 2 labels (ex. *Self-direction: action*, *Security: societal* and *Conformity: rules*), and the concreteness of the 54 labels of human values (ex. *Have privacy*, *Have a safe country*, and *Be compliant*) contribute to the vagueness of 20 labels of level 2 labels by making premises and level 2 labels relatively comparable.

The main contributions of this paper is; (1) the attempt to use similarity score with different levels of labels in a hierarchical dataset to improve the label prediction; (2) the inspection of the effectiveness of SBERT for human value categories detection.

3 System Overview

The Baseline model is a BERT embedding for premises with a classification head to output the 20 categories for human values, as proposed by Kiesel et al. (2022).

3.1 Similarity-aware Models

Our main approach is to concatenate similarity score information between premises and each of 54 human value labels or their example sentences to the baseline model. Figure 1 shows the model architecture. Each similarity score is defined by cosine similarity ranging from -1 to 1. Each similarity score per level 1 label is calculated by SBERT (Reimers and Gurevych, 2019) given a pair of a premise and each level 1 label or its example sentence. Examples of label 1 and example sentences are in table 1. In the model architecture image in figure 1, the left side shows the baseline, and the right side shows the similarity scores between premises and concatenated example sentences of level 1 labels. The concatenation method of each level 1 labels and example sentences are discussed in section 3.2.

Table 1: Several level 1 labels for the same level 2 category (Mirzakhmedova et al., 2023)

Level 2	Level 1	Example Sentences
Self-direction: thought	Be creative	allowing for more creativity or imagination
		being more creative
		fostering creativity
		...
	Be Curious	being the more interesting option
		fostering curiosity
		...
	Have freedom of thought	allowing people to figure things on their own
		..
Self-direction: action	Be choosing your own goals	allowing people to follow their dreams
		...
	...	...

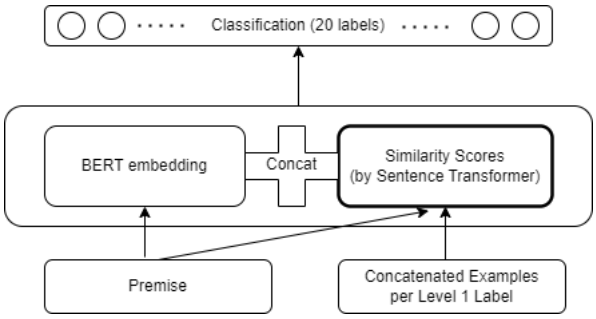

Figure 1: Hierarchical Similarity-aware model Architecture

3.2 Concatenating the same category's values or sentences into a string

In many cases, there are several level 1 labels categorized in the same level 2 labels and several example sentences for the same level 1 label as in table 1. For example, three level 1 labels *Be creative*, *Be curious*, and *Have freedom of thoughts* are categorized in the same level 2 label *Self-direction: thought*.

The same trend applies to example sentences to the level 1 labels. The two sentences *being the more interesting option* and *fostering curiosity* are for the same level 1 label *Be Curious*. Those several labels and sentences have different aspects of the same label for which they are categorized, so we hypothesized all of which should be considered for calculating similarity scores. However, SBERT takes only a sentence to calculate the Similarity Score with premises.

Thus, before calculating the similarity score, we concatenate the level 1 labels that is in the same level 2 category, into one string, or concatenate the example sentences that is in the same level 1 values into one string. For example, for the level 1 label *Be creative*, the example sentences will be concatenated with separator token `</s>`, into one string; `"allowing for more creativity or imaginations </s> being more creative </s> fostering creativity </s> promoting imagination"`.

After the string concatenation, SBERT calculates the similarity score between the premises and each of the concatenated strings. The challenging part of this approach for SBERT will be whether the model can calculate the similarity accurately when one of the inputs is a normal sentence, and the other is a concatenated string with the special tokens and multiple different level 1 labels or example sentences. However, we belief that it will still calculate a reasonable similarity score, since the model is specially trained for taking the semantic similarity of a pair of sentence instead of the syntactic similarity (Reimers and Gurevych, 2019), so that the structural change would not significantly affect the similarity scores.

4 Experimental Setup

We limited our experiments to the main dataset and used the provided data splits as defined in Mirzakhmedova et al. (2023). For system development and hyperparameter tuning, we trained on the train set and evaluated on the validation set. For the final submission, we retrained the model with the best hyperparameters on both train and validation set. The best hyperparameters found can be seen in Appendix A.

For the fine-tuning of SBERT, we combined each premise with each of the 54 concatenated examples representing the level 1 labels (as described in section 3). We assigned a gold label of 1 where the premise draws on the value and a label of -1 where it does not. This is arguably skewing the meaningfulness of the similarity score as -1 generally implies opposite meaning, whereas the different values are often independent, making 0 the better label. We did however find that for our purpose labeling these cases as -1 showed the best overall model improvement. The SBERT model was fine-tuned for 20 epochs on the train set and evaluated on the validation set.

Evaluation was done with label-wise and macro-average F_1-scores, matching both Kiesel et al. (2022) and the official SemEval evaluation. For training, the *transformers*[2] library was used (Wolf et al., 2020). Every model was trained for 20 epochs, evaluated at the end of each epoch and only the best model according to the evaluation metric was kept.

5 Results

For the main quantitative findings, the official results of the SemEval2023 competition are used. They are evaluated on the hidden test set and as such comparable to other teams.

Since, as of writing this paper, the test labels have not yet been published, the analysis sections use the validation split for their evaluation and their findings are thus only comparable between each other.

[2] https://github.com/huggingface/transformers

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
noam-chomsky	.47	.51	.59	.15	.28	.59	.36	.47	.22	.72	.61	.48	.56	.36	.15	.51	.23	.71	.78	.40	.41

Table 2: Official achieved F_1-score of team noam-chomsky in the SemEval2023 competition, 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.

5.1 Main Quantitative Findings

Table 2 shows the official scores of team noam-chomsky on the SemEval2023 competition. This model is consisted of a BERT classifier and a fine-tuned SBERT with concatenated examples as similarity comparisons, hyperparameter-tuned on the validation set and trained on both train and validation set.

With a macro-average F_1-score of 0.47, our model still falls short of the best performing approaches of other teams but does show a sizeable improvement over the BERT baseline.

In particular the classes with the lowest BERT scores show significant gains, with *stimulation* and *humility* more than doubling their previous scores. Only one class, *security: personal*, which also happens to be the best performing class in the BERT baseline, shows a small performance decrease.

Overall we scored middle of the pack at rank 19 out of the 39 participating teams.

5.2 Quantitative Analysis

Since our model is made up of two main parts, it is interesting to look at how each of these parts performs on their own to deduce how they contribute to the final scores in the full model. The scores of these parts on the validation set can be seen in table 3.

The first half of the model effectively matches the baseline model, a simple BERT embedding followed by a classification head. For the second half, the sentence similarity module, we tried a model that performs the classification task only based on the similarity scores. However this model struggled to learn anything, only managing an overall F_1-score of 0.20 with many classes having a score of 0. This suggests that similarity is not able to indicate the classes on its own and instead is just auxiliary information for the main model.

6 Conclusion

Since research on automated human value detection is fairly new, we decided to try an approach that has been shown to work well in other domains to see if similar effects could be observed on this task, We hypothesized that a sentence similarity based approach would do well for a human value detection task as the important information should be less in the individual words and more in a deeper semantic meaning of the sentences.

To this end, we extended the BERT based baseline with a fine-tuned sentence transformer model that compares representations of the premises with those of example sentences for each value. Our results show a significant improvement over the baseline, especially in those categories that the baseline model failed to recognize. Similarity score as an input feature however does not seem to stand on its own and should be considered an auxiliary module only.

Overall this was a successful project to show on another domain that sentence similarity can help detecting deeper meaning that isn't apparent at the surface level.

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
Validation																					
Bert-only	.40	.46	.55	.21	.30	.62	.28	.48	.12	.72	.62	.37	.45	.16	.08	.57	.24	.63	.60	.15	.44
Similarity-only	.20	.38	.50	.00	.00	.53	.00	.15	.00	.67	.63	.00	.41	.00	.00	.03	.00	.61	.00	.01	.04
Final model	.44	.52	.56	.24	.41	.64	.25	.46	.23	.74	.61	.45	.49	.21	.09	.61	.25	.63	.70	.19	.51

Table 3: Macro-average and per class F_1-scores of different model parts on the validation set.

References

Maike Behrendt and Stefan Harmeling. 2021. Argue-BERT: How to improve BERT embeddings for measuring the similarity of arguments. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 28–36, Düsseldorf, Germany. KONVENS 2021 Organizers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gilad Feldman. 2021. Personal values and moral foundations: Examining relations and joint prediction of moral variables. *Social Psychological and Personality Science*, 12(5):676–686.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

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: Identification of human values behind arguments. In *17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

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.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Hyperparameters

Optimizer	AdamW
Batchsize	2
Learning Rate	3e-5
Dropout	0.1
Weight Decay	0.01

B Noam Chomsky

Noam Chomsky, often called "the father of modern linguistics", played an important role in both the field of linguistics and the field of theoretical computer science. As we are an interdisciplinary team with backgrounds in linguistics and computer science, we found he perfectly represents our combination of forces.