# Sina at SemEval-2023 Task 4: A Class-Token Attention-based Model for Human Value Detection

**Omid Ghahroodi**[◇*], **Mohammad Ali Sadraei**[◇*], **Doratossadat Dasgheib**[◇*]
**Mahdieh Soleymani Baghshah**[†], **Mohammad H. Rohban**[†]
, **Hamid R. Rabiee**[†], **Ehsaneddin Asgari**[§]
[◇]DH-NLP Lab, Computer Engineering Department,
Sharif University of Technology, Tehran, Iran.
[†]Computer Engineering Department, Sharif University of Technology, Tehran, Iran.
[§]AI Innovation Center, Data:Lab, Volkswagen AG, Munich, Germany.
{omid.ghahroodi98, m.sadraei, soleymani, rohban, rabiee}@sharif.edu,
d_dastgheib@sbu.ac.ir, asgari@berkeley.edu

## Abstract

The human values expressed in argumentative texts can provide valuable insights into the culture of a society. They can be helpful in various applications such as value-based profiling and ethical analysis. However, one of the first steps in achieving this goal is to detect the category of human value from an argument accurately. This task is challenging due to the lack of data and the need for philosophical inference. It also can be challenging for humans to classify arguments according to their underlying human values. This paper elaborates on our model for the SemEval 2023 Task 4 on human value detection. We propose a class-token attention-based model and evaluate it against baseline models, including finetuned BERT language model and a keyword-based approach.

## 1 Introduction

The social sciences and humanities provide insight into understanding the world and its people, with a primary responsibility of solving human-based issues and providing recommendations. The study of human argumentation and causality is an approach that aids in understanding human relationships and culture, with applications in areas such as faceted search (Amsterdamer and Gáspár, 2022), value-based argument generation (Bostrom et al., 2022), and value-based personality profiling (Liu et al., 2019). The Semantic Evaluation 2023 includes the human value detection task (Kiesel et al., 2023), which aims to classify the human value category based on textual argument.

Our study investigates the effectiveness of keyword extraction and attention-based neural models. Our findings indicate that while context keywords contain the primary argument value, incorporating the class embedding of arguments as queries that focus on the most important concepts of each argument can improve classification results. We also observed that simple models like SVM (Cortes and Vapnik, 1995) perform well compared to neural networks due to the dataset's small size, multi-class prediction, and numerous labels.

We participated in the human value detection task at TIRA (Fröbe et al., 2023) and achieved an average score of 0.47 for all labels, which was 0.09 less than the first team's score. However, we attained the best F1 score of 0.54 for the *Power: Resources* label, which was 0.01 better than the top-performing approach. These findings suggest that there is room for improvement in the task.

To facilitate easier evaluation and reproducibility of our results, we have made our baselines and proposed models[1] available open-source as several Jupyter-notebooks and docker images.

## 2 Background

Human values are ubiquitous in social sciences, and identifying them in argumentative texts can help understand cultures, conflicting beliefs, and opinions. In the human value detection classification task, it is required to determine human values, given human arguments containing premises and conclusions (Kiesel et al., 2022).

### 2.1 Dataset

The task dataset comprises 9324 arguments with corresponding classes from various sources, such as political and religious texts, newspapers, and free-text arguments (Mirzakhmedova et al., 2023). The train, validation, and test sets contain 5393, 1896, and 1576 data items. There is an additional validation and test dataset from community discussion and religious texts. For more information

[*]Equal contribution

[1]`https://github.com/language-ml/human-value-detection`



*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2164–2167
July 13-14, 2023 ©2023 Association for Computational Linguistics

| | train | validation | test |
|---|---|---|---|
| main | 5393 | 1896 | 1576 |
| Nahj | - | - | 279 |
| Zhihu | - | 100 | - |

Table 1: Overview of the available arguments for the detection of the human value

about the dataset, refer to Table 1. The dataset contains three components: premise, stance, and conclusion representing a moral inference. The objective is to determine the value type employed to make this inference. This task is multi-label and multi-class classification.

## 3 System Overview

This section provides a review of the baseline methods and the proposed methods for addressing the task at hand, including keyword extraction and attention-based models.

### 3.1 Baselines

We utilized two baseline models for our experiments: a Support Vector Machine (SVM) and a fully connected neural network. To obtain the sentence embedding, we combined each input sentence's premise, stance, and conclusion parts and fed them to the LABSE model (Feng et al., 2022), which we found to be more appropriate than traditional models like Word2Vec (Mikolov et al., 2013) or BERT (Devlin et al., 2019) for our task.

The fully connected neural network consisted of four layers, and to ensure stable training, we incorporated batch normalization (Ioffe and Szegedy, 2015). We also used dropout (Srivastava et al., 2014) to counter overfitting and improve the model's accuracy. We utilized the sigmoid activation function in the last layer with a threshold of 0.5 to predict the output class. On the validation dataset, this model achieved macro-F1 and micro-F1 scores of 0.32 and 0.47, respectively. The SVM model with a linear kernel and LABSE embedding obtained macro-F1 and micro-F1 scores of 0.31 and 0.46 on the validation dataset, respectively.

Our experiments demonstrated that even simple models such as SVM can perform comparably to neural networks for our task. To further enhance the performance of the SVM, we employed an ensemble of seven SVMs with different kernel functions, including polynomials (with 2, 3, and 4 degrees), RBF, and sigmoid. We randomly selected the kernel function for each SVM to reduce their total variance. This ensemble model achieved macro-F1 and micro-F1 scores of 0.41 and 0.51, respectively, on the validation dataset.

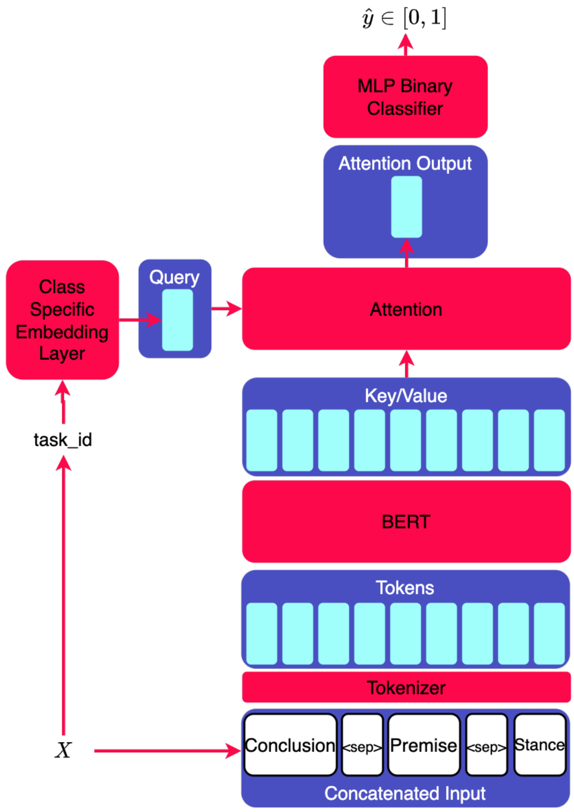


Figure 1: Attention-based model diagram for an example class "Humility"

### 3.2 Keyword Extraction

Since keywords within text data can encompass the fundamental concepts of the text, and these concepts are essential in deriving conclusions from premises, the personal values of the individual making inferences can influence these keywords. Consequently, we developed an approach based on keywords to predict human values. This involved extracting keywords from each human value's class descriptions and training data using Yake (Campos et al., 2020). We then assigned positive labels to data classes with scores above a pre-defined threshold based on the number of intersections between arguments in the test data and each class's keywords. We also repeated this approach using the embeddings of keywords instead of the surface forms. Despite these efforts, both experiments yielded poor results, with an average F1 score overall categories of only 0.28, which was only slightly better than the 1-Baseline provided by organizers.

| 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 |
| Attention-based approach | .47 | .42 | .60 | .20 | .21 | .62 | .39 | .54 | .24 | .74 | .58 | .46 | .51 | .52 | .19 | .50 | .24 | .71 | .78 | .36 | .49 |
| *Nahj al-Balagha* | | | | | | | | | | | | | | | | | | | | | |
| Best per category | .48 | .18 | .49 | .50 | .67 | .66 | .29 | .33 | .62 | .51 | .37 | .55 | .36 | .27 | .33 | .41 | .38 | .33 | .67 | .20 | .44 |
| Best approach | .40 | .13 | .49 | .40 | .50 | .65 | .25 | .00 | .58 | .50 | .30 | .51 | .28 | .24 | .29 | .33 | .38 | .26 | .67 | .00 | .36 |
| BERT | .28 | .14 | .09 | .00 | .67 | .41 | .00 | .00 | .28 | .28 | .23 | .38 | .18 | .15 | .17 | .35 | .22 | .21 | .00 | .20 | .35 |
| 1-Baseline | .13 | .04 | .09 | .01 | .03 | .41 | .04 | .03 | .23 | .38 | .06 | .18 | .13 | .06 | .13 | .17 | .12 | .12 | .01 | .04 | .14 |
| Attention-based approach | .25 | .07 | .21 | .00 | .40 | .60 | .12 | .00 | .12 | .38 | .19 | .26 | .22 | .17 | .22 | .28 | .18 | .22 | .29 | .12 | .27 |
| *New York Times* | | | | | | | | | | | | | | | | | | | | | |
| Best per category | .47 | .50 | .22 | - | .03 | .54 | .40 | - | .50 | .59 | .52 | - | .33 | 1.0 | .57 | .33 | .40 | .62 | 1.0 | .03 | .46 |
| Best approach | .34 | .22 | .22 | - | .00 | .48 | .40 | - | .00 | .53 | .44 | - | .18 | 1.0 | .20 | .12 | .29 | .55 | .33 | .00 | .36 |
| BERT | .24 | .00 | .00 | - | .00 | .29 | .00 | - | .00 | .53 | .43 | - | .00 | .00 | .57 | .26 | .27 | .36 | .50 | .00 | .32 |
| 1-Baseline | .15 | .05 | .03 | - | .03 | .28 | .03 | - | .05 | .51 | .20 | - | .07 | .03 | .12 | .12 | .26 | .24 | .03 | .03 | .33 |
| Attention-based approach | .24 | .11 | .00 | - | .00 | .29 | .00 | - | .33 | .57 | .31 | - | .23 | .67 | .00 | .21 | .31 | .27 | .33 | .00 | .38 |

Table 2: Achieved $F_1$-score of team Sina (Seyyed Hossein Nasr) 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.

Therefore, we can conclude that although the main concepts of argumentative texts may contain informative data about human values, there are strong hidden connections between keywords and human values that result in the human values not automatically discernibly classifiable.

### 3.3 Attention-based Model

In this model, we utilized BERT to compute embeddings for every token in the concatenated input. Subsequently, we implemented an attention layer over these embeddings to generate a singular embedding. This attention layer utilized the token embeddings as both the value and key, while a class-specific embedding served as the query. These class-specific embeddings were randomly initialized and then learned during the training phase. The final step involved using a binary classifier to predict if the input belonged to the selected class. The entire network was trained end-to-end without freezing the BERT model. We provided a visual representation of this architecture in Figure 1.

To address the issue of imbalanced training data, where most of the labels were negative, we oversampled the positive data to achieve balance in each epoch. Our experiments showed that without this technique, the models did not converge.

## 4 Results

Table 2 presents the results obtained by applying the method described in section 3.3. We compare it with the best approach, baselines provided by the organizers, and the best model per category. The overall average F1 score for 20 labels was 0.47, which outperformed the organizers' 1-Baseline and BERT models. Our difference with the best approach was 0.09, and we even surpassed it by 0.01 in the F1 score of the *power: Resources* label. These results show that using class-specific embedding can be effective in any attention-based approach.

For Nahj al-Balagha data, the result of our model was better than the 1-Baseline; it could not improve the BERT baseline, and to improve the result, we need more related human-value data from a religious source.

## 5 Conclusion

Detecting human values is a practical task that challenges natural language processing methods because of the necessity to comprehend moral and philosophical concepts. This paper proposes a solution to this problem by learning class-specific embedding and utilizing an attention mechanism to find the best features for a binary classifier. Although we have introduced novel models to address this issue, we have not achieved the desired level of accuracy because we used BERT-base instead of models with more parameters and a single transformer-based model due to our computational resource constraints. Given the paucity of data for numerous classes, unsupervised techniques such as training a dedicated language model for philosophical and moral text could be utilized in future research.

## References


Yael Amsterdamer and Laura Gáspár. 2022. Interactive knowledge graph querying through examples and facets. In *New Trends in Database and Information Systems*, pages 201–211, Cham. Springer International Publishing.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural Lnagugae Dedution through Searcg over Statement Compositions.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

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)*.

Hui Liu, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu, and Weijun Wang. 2019. Personality or value: A comparative study of psychographic segmentation based on an online review enhanced recommender system. *Applied Sciences*, 9(10).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.