# CSECU-DSG at SemEval-2023 Task 4: Fine-tuning DeBERTa Transformer Model with Cross-fold Training and Multi-sample Dropout for Human Values Identification

**Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy**
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

## Abstract

Human values identification from a set of argument is becoming a prominent area of research in argument mining. Among some options, values convey what may be the most desirable and widely accepted answer. The diversity of human beliefs, random texture, and implicit meaning within the arguments make it more difficult to identify human values from the arguments. To address these challenges, SemEval-2023 Task 4 introduced a shared task ValueEval focusing on identifying human value categories based on given arguments. This paper presents our participation in this task where we propose a fine-tuned DeBERTa transformers-based classification approach to identify the desired human value category. We utilize different training strategies with the fine-tuned DeBERTa model to enhance contextual representation on this downstream task. Our proposed method achieved competitive performance among the participants' methods.

## 1 Introduction

In human decision-making, values play a crucial role. The meaning of values often refers to any individual's personal beliefs and priorities about any argument which tends to go to the conclusion of that argument (Teze et al., 2019). The conclusion may vary from person to person though the arguments are the same. Identifying human value categories which drive a human being to conclude an argument and make perspective decisions is very crucial to build a value-driven artificial intelligence (AI) agent to mitigate future security risks from those AI agents. Nevertheless, it is beneficial for various NLP tasks also including argument mining, sentiment analysis, political discourse analysis, personality analysis, and marketing analysis.

However, identifying human values from the arguments is the most challenging and prominent task in NLP. Because of the large variety and heterogeneity of personal beliefs which may be ambivalent for the same group of people regarding their personal experiences, this task becomes challenging. To address the challenges of human values identification from arguments Kiesel et al. (2023) introduce a shared task ValueEval at SemEval-2023 [1]. The task is composed of three categories of tracks including the argument test, Nahj al-Balagha test, and New York Times test datasets. Here, the argument test considers the main dataset for this task. To categorise the arguments, this task articulates 20 fixed categories of values which are shown in Figure 1.

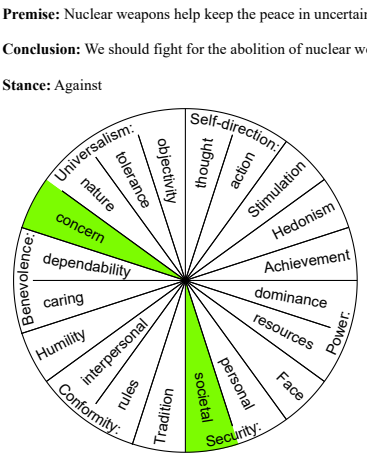


Figure 1: Example and categories of the human value of ValuEval shared tasks. Here highlighted categories indicate the desired value categories of the example.

The task introduces the relation between a premise argument and a conclusion argument where the stance conveys that the conclusion is in favor of or against the premise. In the following example, *"nuclear weapons help keep the peace in uncertain times"* the argument is considered a premise and the conclusion is *"We should fight for the abolition of nuclear weapons"*, which is a stance against

**The first two authors have equal contributions.

[1] https://touche.webis.de/semeval23/touche23-web/index.html



*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1988–1994

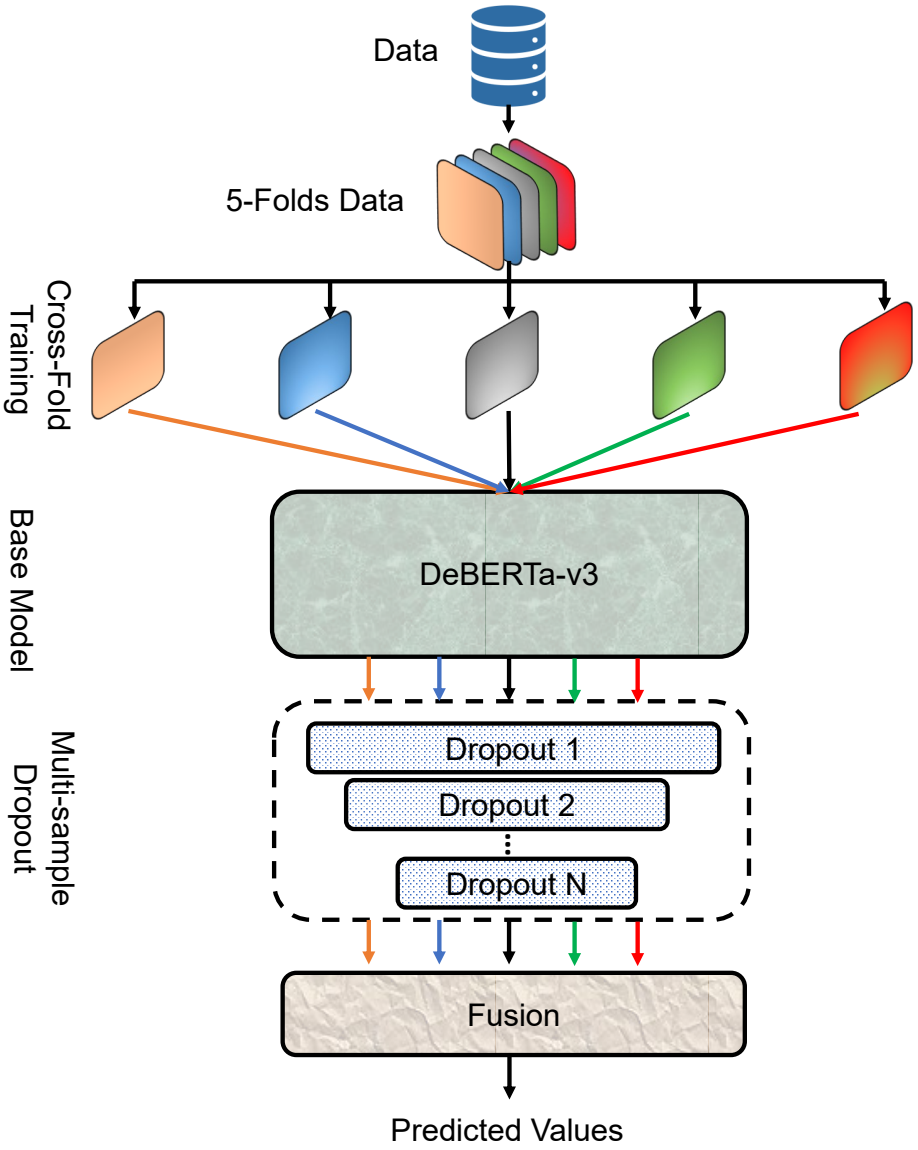


Figure 2: Overview of proposed method

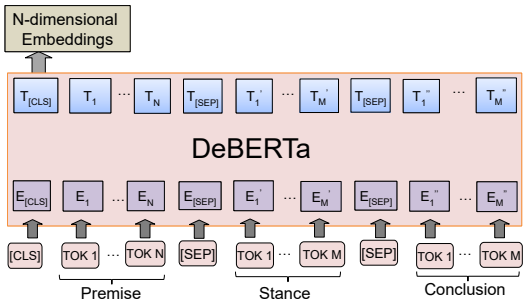


Figure 3: Input representation of our proposed method.

the premise tenor since many people agree that nuclear weapons have the potential to damage human civilization. In spite of this, the dataset assumes that the conclusion stance is against the premise for the reasons of Security: societal, and Universalism: concern values.

Prior works (Kiesel and Alshomary, 2022; Handke and Stein) of identifying human values based on arguments only explored theoretical and manual aspects but never attempted to build an automated classification system. However, we proposed a DeBERTa-v3 transformer-based automated classification system in this paper to classify human values based on three parameters premise, conclusion, and stance. We also used different training strategies including cross-fold training and multi-sample dropout training strategies to train our proposed method.

Accordingly, the remaining sections of the paper are organized as follows: Section 2 introduces our proposed system for automatically identifying human value from given arguments, while Section 3

presents our system design and parameter settings, which is followed by Section 4 where we discuss our results, and performance analysis. Finally, we conclude with some future directions in Section 5.

## 2 Human Values Identification Framework

In this section, we describe our proposed human values identification framework. This task is designed to be a multi-label classification task. The purpose of this study is to determine what human values are present in the argument provided. A brief overview of the proposed framework we have developed in this work is depicted in Figure 2.

In our proposed framework, we use a sentence pair training concept in the transformer models in order to perform human values identification. The idea behind this concept is to pack together the input premise, stance, and conclusion into one sequence as shown in Figure 3. In our approach, we employ a pre-trained transformer model, DeBERTa, in order to extract contextual features from packed sequences based on given contexts. The DeBERTa model has been fine-tuned for the human value identification task so that it captures domain-specific information about human values within the context. In addition to the extracted feature vectors, we later use a multi-sample dropout (Inoue, 2019; Du et al., 2022) procedure to improve the generalization ability of the system output. To predict class confidence, a classification head averages the feature vectors from the multi-sample dropout. A cross-fold training technique reduces the error rates associated with class label prediction (Reul et al., 2018). In order to improve prediction performance, we employ a 5-cross-fold training in our method. Finally, for the effective fusion of the class confidences, we take the arithmetic mean of the prediction scores obtained from each trained model for each fold.

### 2.1 Transformers Model

Unlike other models, transformer models are capable of distilling long-term dependencies and improving the relationship between the words in a sentence by incorporating multi-head attention and positional embedding mechanisms. In order to obtain the contextualized features representation of argument context, we fine-tuned the DeBERTa transformer model.

#### 2.1.1 DeBERTa

DeBERTa (He et al., 2020) stands for decoding enhanced BERT with disentangled attention. Using a disentangled attention mechanism and an enhanced mask decoder, it improves the BERT and RoBERTa models. This study utilized the enhanced version of the DeBERTa model, known as DeBERTa-V3 (He et al., 2021). The DeBERTa-V3 model utilized the ELECTRA style pre-training. In DeBERTa-V3, the ELECTRA model's mask language modeling (MLM) has been replaced with a replaced token detection (RTD) strategy. In order to determine whether an input token is original or if it has been replaced by a generator, the model is trained as a discriminator. In addition, it implements the gradient-disentangled embedding sharing (GDES) method, which allows embeddings to be shared between generators and discriminators. As a result of this unidirectional sharing, the generator shares its embeddings with the discriminator, but the discriminator has the ability to only backpropagate the embeddings. With the above component changes, the DeBERTa model has been able to achieve significant improvements in many downstream tasks. Motivated by this, we intend to extract the feature representation of the arguments by using Huggingface's (Wolf et al., 2019) implementation of the *microsoft/deberta-v3-large* checkpoint [2]. In the embedding layer, it consists of 24 transformer blocks, a hidden size of 1024, and 131M parameters as well as a vocabulary of 128K tokens.

### 2.2 Different Training Techniques

Several studies (Du et al., 2022) have demonstrated that the transformer model performance can be improved using various training strategies. In order to achieve this, we utilized two different training strategies, including multi-sample dropouts and 5-cross-fold training.

#### 2.2.1 Cross-fold Training

We employ a stratified cross-fold training strategy (Sechidis et al., 2011; Pikrakis and Theodoridis, 2014; Reul et al., 2018) in our proposed method to reduce the error rates during the model training process, thus improving the robustness of our model. By taking samples from these disjoint groups, it preserves the proportion of disjoint groups within a population. As opposed to training a model with the full dataset, this method basically uses the train-

[2] https://huggingface.co/microsoft/deberta-v3-large

ing sample to create a bunch of folds and each fold is then used to train the model with each fold dataset. As a result, it makes a significant contribution to the tuning of hyperparameters and accordingly captures the diversity of contexts associated with this task effectively. In our method, we utilize five-fold stratified multi-label cross-fold training.

#### 2.2.2 Multi-sample Dropout

In deep neural networks, dropout is one of the most efficient regularization instruments for preventing overfitting. It works by randomly discarding neurons during training; as a result, generalization occurs because neurons no longer depend on one another. The multi-sample dropout-based training strategy (Inoue, 2019) increases generalization ability and speeds up the training of the base model over the original dropout, which in turn enhances the overall performance of the system. Using this training strategy, we apply five dropout samples to our transformer-based model. Following the dropout layer, we duplicate the features vector of the transformer model, sharing the weights between the five duplicated fully connected layers. A final loss can be obtained by aggregating all the losses observed in each sample and taking the average of all these losses.

## 3 Experimental Setup

In this section, we now describe the dataset and hyper-parameters settings we have used in our proposed method in brief.

### 3.1 Dataset Description

To evaluate the performance of participants' systems at the ValueEval-2023 shared task (Kiesel et al., 2023) in SemEval-2023, the organizers utilized a benchmark dataset (Mirzakhmedova et al., 2023) based on Webis-ArgValues-22 (Kiesel et al., 2022) published in ACL-2022. Table 1 summarizes the statistics of the dataset. The dataset comprises a cross-cultural value taxonomy including Africa, China, India, and USA cultures. Each sentence is annotated with 54 values in 20 categories including *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, and universalism: objectivity*. Organizers provide three subsets of test data including arguments-test, test-nahjalbalagha (Nahj al-Balagha), and test-nyt (New York Times news). However, arguments-test is the base test set of ValueEval-2023.

| Category | Data |
|---|---|
| Train | 5393 |
| Validation-zhihu | 100 |
| Arguments-test | 1576 |
| Test-nahjalbalagha | 279 |
| Test-nyt | 80 |
| Total | 7428 |

Table 1: The statistics of used dataset in ValuEval-2023 shared task.

### 3.2 Experimental Setting

We are now going to present a detailed description of our experimental settings and the hyper-parameter settings that we have applied in the fine-tuning strategy to design our proposed system to be used for the ValueEval 2023 shared task (Kiesel et al., 2023). We fine-tune a state-of-the-art Huggingface (Wolf et al., 2019) transformers model named DeBERTa for this task. For reproducible results, we used a CUDA-supported GPU and set the manual seed equal to four. Table 4 in the appendix, we present the optimal parameter settings for our proposed model based on the validation data. We have used AdamW optimizer (Kingma and Ba, 2014) with a weight decay. The maximum patience number of early stopping is set to 5. We use the dropout range of 0.1 to 0.5, in the multi-sample dropout training. Later, in our 5-cross-fold training phase, we combine the training and development data for efficient training.

### 3.3 Evaluation Measure

The ValueEval 2023 shared task organizers employed various standard evaluation metrics including the F1-score, precision, and recall to evaluate the participants' systems. However, averaging over all value categories, the averaged F1-score is considered the primary evaluation measure of this task.

## 4 Experimental Results and Analysis

### 4.1 Overall Result

In this section, we analyze the comparative performance of our proposed CSECU-DSG (team fazlur-

| 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 |
| CSECU-DSG | .49 | .54 | .69 | .12 | .26 | .60 | .32 | .48 | .02 | .77 | .66 | .64 | .53 | .29 | .08 | .55 | .28 | .78 | .82 | .37 | .51 |
| Adam Smith | .56 | .59 | .71 | .22 | .29 | .66 | .48 | .52 | .30 | .79 | .67 | .65 | .61 | .61 | .19 | .60 | .36 | .74 | .84 | .41 | .53 |
| John Arthur | .55 | .56 | .70 | .27 | .25 | .65 | .50 | .52 | .39 | .76 | .60 | .63 | .60 | .69 | .24 | .55 | .41 | .74 | .86 | .44 | .58 |
| Stanley Grenz | .50 | .55 | .67 | .10 | .29 | .61 | .34 | .49 | .18 | .77 | .65 | .62 | .52 | .29 | .12 | .57 | .23 | .75 | .79 | .38 | .42 |
| Prodicus | .48 | .53 | .61 | .07 | .27 | .54 | .32 | .41 | .15 | .73 | .62 | .54 | .51 | .35 | .11 | .53 | .15 | .73 | .78 | .37 | .43 |
| Noam Chomsky | .47 | .51 | .59 | .15 | .28 | .59 | .36 | .47 | .22 | .72 | .61 | .48 | .56 | .36 | .15 | .51 | .23 | .71 | .78 | .40 | .41 |
| Quintilian | .38 | .49 | .58 | .00 | .00 | .58 | .23 | .44 | .00 | .66 | .52 | .47 | .49 | .00 | .00 | .41 | .30 | .65 | .64 | .38 | .45 |
| *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 |
| CSECU-DSG | .29 | .15 | .29 | .00 | .40 | .57 | .00 | .00 | .00 | .41 | .24 | .00 | .25 | .00 | .09 | .36 | .38 | .28 | .67 | .00 | .28 |
| Adam Smith | .40 | .13 | .49 | .40 | .50 | .65 | .25 | .00 | .58 | .50 | .30 | .51 | .28 | .24 | .29 | .33 | .38 | .26 | .67 | .00 | .36 |
| Stanley Grenz | .35 | .16 | .39 | .00 | .36 | .64 | .25 | .00 | .30 | .48 | .26 | .40 | .28 | .22 | .22 | .36 | .25 | .33 | .67 | .00 | .44 |
| Prodicus | .30 | .17 | .33 | .00 | .40 | .59 | .00 | .00 | .37 | .42 | .27 | .53 | .26 | .07 | .00 | .38 | .35 | .23 | .00 | .17 | .41 |
| Noam Chomsky | .26 | .09 | .14 | .00 | .44 | .54 | .10 | .13 | .24 | .50 | .19 | .42 | .30 | .13 | .13 | .34 | .22 | .21 | .20 | .11 | .32 |
| Quintilian | .20 | .03 | .16 | .00 | .10 | .46 | .13 | .20 | .14 | .38 | .19 | .49 | .19 | .00 | .24 | .25 | .16 | .18 | .00 | .00 | .22 |

Table 2: $F_1$-score of team fazlur-rahman (CSECU-DSG) per test dataset along with other selected participant's method and baselines, 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.

rahman) system in the ValueEval-20223 (Kiesel et al., 2023) human values identification shared task. The comparative performance of our proposed CSECU-DSG system on test data against other participants' systems and baseline systems are presented in Table 2.

In ValueEval-2023, we made submissions for the argument-test and test-nahjalbalagha test subset. Initially, we presented the performance of the highest-performing participant approach for each individual category, the best participant approach, as well as the organizer's BERT and 1-Baseline approaches. Then, we present our proposed CSECU-DSG (team fazlur-rahman) method performance based on F1-score. We also presented the performance of selected participating systems in ValueEval-2023. Based on the primary evaluation metric averaged F1-score on argument-test and test-nahjalbalagha test subsets, the best-performing system, Adam Smith achieved 0.56 and 0.40, respectively. Our proposed system obtained 0.49 and 0.29 on the argument-test and test-nahjalbalagha test subsets, respectively. According to this analysis, we can observe that our proposed method achieved competitive performance in comparison with the other selected systems and baseline approaches. This deduces the applicability and generalizability of our system for the human values identification task.

### 4.2 Discussion

We have performed various ablation studies on different folds cross-fold training strategies and the performances are shown in Table 3. We employ 5-cross-fold training for our proposed CSECU-DSG

| Method | Average F1-score |
|---|---|
| CSECU-DSG (5-Fold) | 0.49 |
| *Performance of different folds cross-fold training* | |
| 3-Fold | 0.45 |
| 2-Fold | 0.13 |

Table 3: Performance analysis of different folds cross-fold training against the argument-test dataset.

method which achieved a 0.49 averaged F1-score on the base test subset whereas the ablation study shows that 3-fold and 2-fold scored 0.45 and 0.13 averaged F1-score respectively. These results deduce the effectiveness of our 5-cross-fold training strategies. Because of too many labels, lower cross-fold training drastically failed to acquire complex pattern knowledge since the labels of the dataset are also imbalanced. It is the reason when we reduce the number of folds the results is also dropped. However, from 3-fold to 2-fold results are dropped by a large margin cause in 2-fold training similar kinds of tokens are overlapping multiple times which may be hampered model learning.

## 5 Conclusion

This paper presents our approach to the human values identification task. We tackled the problem using sentence pair training of a transformer model. We fine-tuned the DeBERTa transformer model with various training techniques including cross-fold training and multi-sample dropout. By using pairwise learning, we exploited the contextual relation between three texts in order to estimate human value categories. Experimental results demonstrated the efficacy of our DeBERTa-based proposed method. Our method achieved competitive performance compared to other participants.

In the future, we will implement other SOTA transformer models and also combine multiple transformer models into a unified architecture. Weighted average fusion could capture better arguments for all human values classes since the dataset is imbalanced.

## References


Xiyang Du, Dou Hu, Jin Zhi, Lianxin Jiang, and Xiaofeng Shi. 2022. Pali-nlp at semeval-2022 task 6: isarcasmeval-fine-tuning the pre-trained model for detecting intended sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 815–819.

Johannes Kiesel Milad Alshomary Nicolas Handke and Xiaoni Cai Henning Wachsmuth Benno Stein. Identifying the human values behind arguments.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Johannes Kiesel and Milad Alshomary. 2022. Identifying the human values behind arguments. In *Annual Meeting of the Association for Computational Linguistics*.

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)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Aggelos Pikrakis and Sergios Theodoridis. 2014. Speech-music discrimination: A deep learning perspective. In *2014 22nd European signal processing conference (EUSIPCO)*, pages 616–620. IEEE.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving ocr accuracy on early printed books by utilizing cross fold training and voting. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 423–428. IEEE.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Juan Carlos Teze, Antoni Perello-Moragues, Lluís Godo, and Pablo Noriega. 2019. Practical reasoning using values: an argumentative approach based on a hierarchy of values. *Annals of Mathematics and Artificial Intelligence*, 87:293 – 319.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

# A Appendix

Hyper-parameters and their optimal value of our proposed method.

| Hyper-parameter | Optimal Value |
|---|---|
| Learning rate | 3e-5 |
| Max-len | 128 |
| Number of epochs | 5 |
| Batch size | 2 |
| Manual seed | 4 |
| Number of fold | 5 |
| weight decay | 0.01 |
| Dropout | 0.1, 0.2,..., 0.5 |

Table 4: Model settings for ValueEval-2022 shared task.