PAI at SemEval-2023 Task 4: A general multi-label classification system with class-balanced loss function and ensemble module.

{malong633,sunzeye273,jiangjiawei358,lixuan208}@pingan.com.cn
Ping An Life Insurance Company of China, Ltd.

Abstract

The Human Value Detection shared task(Kiesel et al., 2023) aims to classify whether or not the argument draws on a set of 20 value categories, given a textual argument. This is a difficult task as the discrimination of human values behind arguments is often implicit. Moreover, the number of label categories can be up to 20 and the distribution of data is highly imbalanced. To address these issues, we employ a multi-label classification model and utilize a class-balanced loss function. Our system wins 5 first places, 2 second places, and 6 third places out of 20 categories of the Human Value Detection shared task, and our overall average score of 0.54 also places third. The code is publicly available at https://www.github.com/diqiuzhuanzhuan/semeval2023.

1 Introduction

The Human Value Detection shared task(Kiesel et al., 2023) aims at building a system to identify human values behind an textual argument. Though human values have been considered in formal argumentation for about 20 years(Bench-Capon, 2003), it is the first task to draw values from arguments computationally. The organizers provide the first dataset Touché23-ValueEval Dataset(Mirzakhmedova et al., 2023) for studying human values behind arguments. It contains 5270 arguments extracted from four different cultural backgrounds: Africa, China, India and USA. All arguments are written in English. Every argument consists of three parts: a premise, a conclusion and a stance indicating whether the premise is in favor of or against the conclusion. Table 1 demonstrates examples from the dataset.

In this paper, we propose a general multi-label classification system for the Human Value Detection shared task. To handle imbalanced data, we apply a class-balanced loss function and perform data augmentation for classes with less data. Transformer-based models have demonstrated remarkable performance across a variety of NLP tasks in recent years. As a result, we adopt these models as encoder in our work. To select the best encoder model, we evaluate 4 candidates, including BERT(Devlin et al., 2018), ALBERT(Lan et al., 2019), RoBERTa(Liu et al., 2019), DeBERTa(He et al., 2020) and etc. The sentence vector computed by the encoder is used as input for the multi-label classifier. We select Sigmoid function as the activation function for the classifier. Given the highly imbalanced nature of the data, we experiment with 5 loss functions that are specifically designed for such case, including focal loss, distribution-balanced loss and etc., and select CB-NTR loss(Huang et al., 2021) as our final loss function. In terms of data preprocessing, we conduct rewriting and data augmentation to solve the weaknesses in the data and the model. In the end, we conduct 8-fold cross-validation and use a weighted voting ensemble based on evaluation scores of each class to get the final prediction.

Besides the system description, we also have some interesting findings: (1) setting individual threshold for each classifier in a multi-label classification model may not be an effective approach, (2) concatenating data that does not conform to human language can have a negative impact on model performance.

2 Background

Despite many years of research on the relationship between human values and arguments, identifying human values from argumentative texts computationally is a novel approach. In this task, the heart of value-based argumentation framework has been determined as a value taxonomy(or a set of values) that is both accepted and relevant by the organizers. The organizers extend the refined theory of (Schwartz et al., 2012) by adding and integrating nine values and building multi-level taxonomy in

conclusion	stance	premise
We should ban human cloning	in favor of	we should ban human cloning as it will only cause huge issues when you have a bunch of the same humans running around all acting the same.
We should legalize cannabis	against	if you start legalizing drugs it could open up the floodgates for more legalization of dangerous drugs.

Table 1: examples of argument.

line with psychological research and publish the Touché23-ValueEval Dataset, in which there are a total of 54 values and 20 value categories. However, in this task, we only care about the classification of 20 value categories. It's worth mentioning that the organizers have already conduct experiments on Touché23-ValueEval Dataset and obtain results with an average F1 score of 0.25.

3 Our System

In this section, we will illustrate how our system works. We use Transformer-based model as our encoder model. We build four kinds of data modules: BaselineArgumentDataset, PremiseArgumentDataset, RewriteArgumentDataset and AugmentArgumentDataset, apply the best loss function, and finally integrate the results from multiple data modules. We also provide an overview of the system in Figure 1.

3.1 Data Module

Based on the characteristics of these data, we construct four different dataset.

BaselineArgumentDataset As mentions in previous section, each argument contains a premise, a conclusion, and a stance. We simply concatenate the three part with a special separator token [SEP], make it work as our baseline data set.

PremiseArgumentDataset As shown in Table 1, the premise field obviously holds the richest information. Only using the premise part is also a good option, so we build a new data module(named after PremiseArgumentDataset) that only uses the premise field.

RewriteArgumentDataset The simple way of concatenating data in BaselineArgumentDataset is not in line with natural language, and therefore is also inconsistent with the input data used by pretrained models. We rewrite all the arguments by modifying the positive or negative attitude in the conclusion and use the premise and the modified conclusion to build new input data. We also provide an illustration in Figure 2.

AugmentArgumentDataset The label distribution is highly imbalanced, to mitigate data imbalance issue, we concatenate ordinary samples with sparse-label samples to augment the sparse-label samples. Concretely, for each sparse sample, we randomly sample from the remaining samples, perform data concatenation, and merge the label of this two samples. The iteration continues until the number of samples in each category can reach 15% of the overall proportion.

3.2 Multi-label Classification Module

We approach Human Value Detection task as a multi-label classification problem. We employ a transformer-based model as our encoder and derive a representation vector $V \in \mathcal{R}^H$ from the [CLS] token's embedding in the output of the encoder. Then we learn $|L|$(the number of labels) binary classifiers $H_l : X \rightarrow \{l, \neg l\}$, one for each different label l in L. We use the Sigmoid function as the activation function and set the activation threshold to 0.5.

$$\bar{y}^i = sigmoid(W_i^T * V + b_i) \quad (1)$$

$$y^i \in l_i \quad if \quad \{\bar{y}^i < 0.5\}$$

$$y^i \in \neg l_i \quad if \quad \{\bar{y}^i \geq 0.5\}$$

where $W^T \in \mathcal{R}^{H*|L|}$ denotes a logic matrix, b denotes a bias, and i is the index of the matrix and the bias.

We also attempt to set different classification thresholds for each label based on its frequency but find that it does not improve the performance.

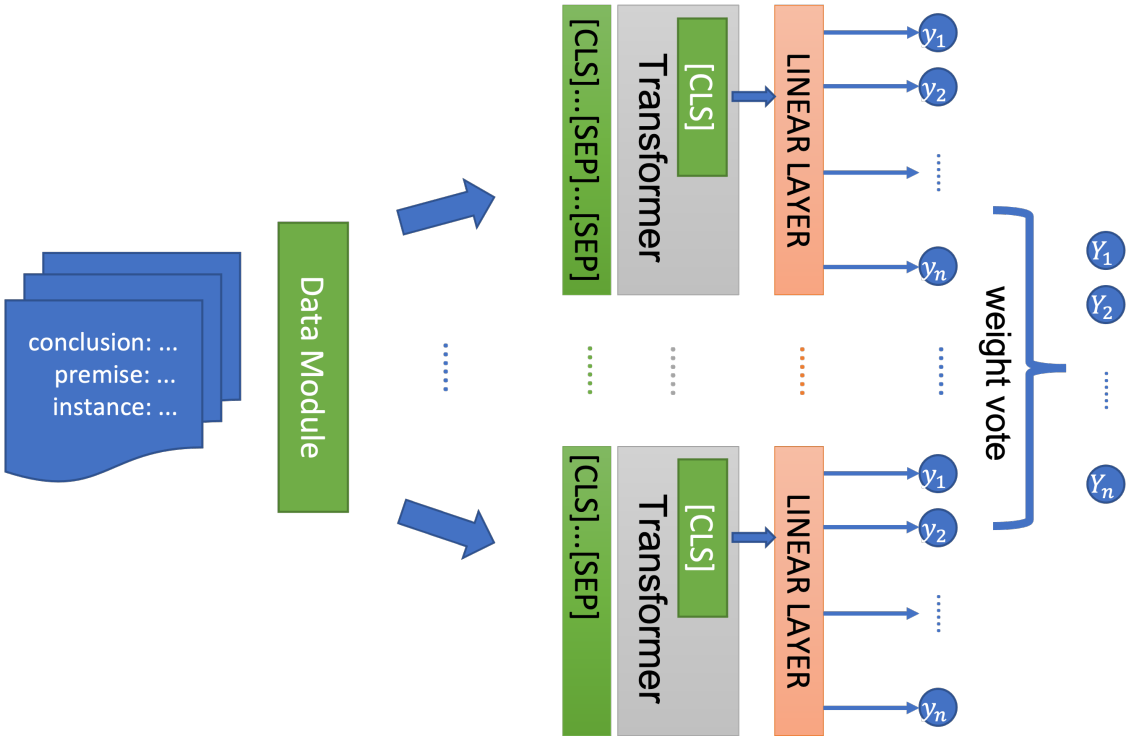

Figure 1: The overall architecture of our proposed system. Data Module generates several different kinds of input data, then feds them into a multi-label classification model individually. In the end, the ensemble module obtains results by weight voting.

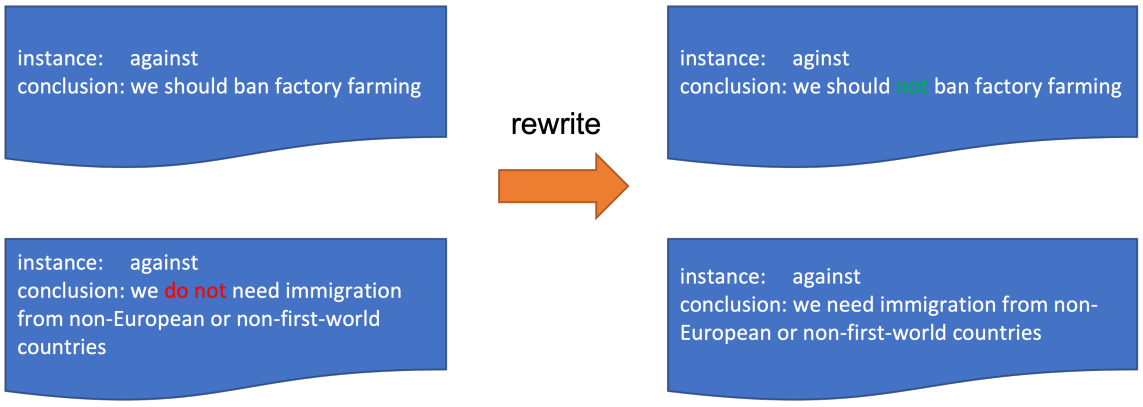

Figure 2: We modify the content of conclusion by inserting or removing some words.

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
2023-01-11-08-14-46	.54	.58	**.71**	.35	.34	.62	.43	.49	.31	**.80**	**.68**	.55	**.61**	.56	.16	.57	.34	.77	.83	.40	.42
2023-01-27-04-05-18	.54	.59	**.71**	.29	.32	.61	.45	.49	.36	.79	.67	.55	.59	.58	.12	.58	.34	.76	.85	.42	.48

Table 2: Achieved F_1-score of our team PAI(theodor-zwinger) over test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. Approaches marked with * were not part of the official evaluation. 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.

3.3 Loss function

Binary Cross Entropy(BCE) (Bengio et al., 2013), widely used for multi-label classification, is our baseline loss function. The main challenge lies in the imbalance of label distribution and the complex dependencies between labels. A lot of work has been published on the design of loss function. Focal loss(FL)(Lin et al., 2017) places a higher weight on samples with low probability to emerge. Class-balanced loss(CB)(Cui et al., 2019) design a re-weighting scheme that uses the effective number of samples for each class. Distributed-balanced loss(DB)(Wu et al., 2020) takes into account the impact caused by label co-occurrence and adds a negative tolerant regularization(NTR) to mitigate the over-suppression of negative labels. CB-NTR loss(Huang et al., 2021) integrates NTR with CB.

We experiment with all the mentioned loss functions and find that the CB-NTR loss function performes the best in this task.

3.4 Ensemble

When submitting the results of the final test set, we apply the model ensemble. Given the m predictions $\{y_{\theta 1}, ..., y_{\theta 3}\}$ generated by m models, following (Ma et al., 2022), we form the final results by voting based on the F1 score of each value category on validation set.

4 Experimental Setup

The implementation is based on PyTorch Lightning[1], Transformers[2], and AllenNLP[3]. During training, we set the batch size as 16, random seed as 42, learning rate as 1e-5, warmup steps as 1000. We use the AdamW optimizer(Loshchilov and Hutter, 2017) and the polynomial decay scheduler with the power of 0.01. We set a maximum epoch of 25 and set patience of 3 for early stop. All experiments are run on one single Tesla V100 GPU.

The macro F1 is our metric. As for data splits, we directly use the official train and development set. But for the test set, we conduct 8-fold cross-validation and ensemble 8 results to one.

5 Results and Analysis

In this section, we display our results and analyze the impact of each component through ablation studies.

5.1 Results

In this competition with a total of 47 teams, we achieve the third place with a macro F1 score of 0.54. We outperform the official baseline by 0.28 and official BERT by 0.12. The result is shown in Table 2.

[1]https://www.pytorchlightning.ai
[2]https://huggingface.co/docs/transformers/index
[3]https://allenai.org/allennlp

Prerained-model	dev macro F1
BERT$_{base}$	0.362
BERT$_{large}$	0.423
RoBERTa$_{base}$	0.422
RoBERTa$_{large}$	0.460
DeBERTa$_{v3_base}$	0.395
DeBERTa$_{v3_large}$	0.443
ALBERT$_{base}$	0.332
ALBERT$_{large}$	

Table 3: Dev macro F1 score of different Transformer-based models.

data module	dev macro F1
BaselineArgumentDataset	0.460
PremiseArgumentDataset	0.462
RewriteArgumentDataset	0.468
AugmentArgumentDataset	0.463

Table 4: Dev macro F1 score of different data modules.

5.2 Ablation Study

We also conduct ablation experiments to validate our designs, including encoder model, data modules, loss function, and ensemble. Due to limited space, we only compare the macro F1 of all categories.

Encoder Model We build our baseline model with BaselineArgumentDataset and BCE loss and run experiments to find out the best encoder model among RoBERTa(Liu et al., 2019), BERT(Devlin et al., 2018), ALBERT(Lan et al., 2019), DeBERTa(He et al., 2020), etc. As shown in Table 3, the large version of RoBERTa achieves the best score.

Data Module We use the best encoder model to conduct ablation experiments on the data module. The results are shown in Table 4

Loss Function We further evaluate different loss functions by average F1 scores, including BCE(Bench-Capon, 2003), FL(Lin et al., 2017), CB(Cui et al., 2019), DB(Wu et al., 2020), and CB-NTR(Huang et al., 2021). As shown in Table 5, CB-NTR loss achieves the best score.

Ensemble Our ensemble approach can significantly improve performance. We integrate the results of different data modules based on their performance on the development and test set. The result is shown in Table 6, where we can see our ensemble approach outperforms the best single-model by 0.04 F1 score over test set.

loss	dev macro F1
BCE	0.460
FL	0.469
CB	0.478
DB	0.469
CB-NTR	0.492

Table 5: Dev macro F1 score of different loss functions.

approach	dev macro F1	test macro F1
no ensemble	0.492	0.498
ensemble	0.512	0.538

Table 6: macro F1 score over dev and test data with ensemble or no ensemble

6 Conclusion

In this paper, we describe our system on how to use multi-label classification model to handle Human Value Detection task(Kiesel et al., 2023). We win 5 first places, 2 second places, and 6 third places on the leaderboard. We analyze both the data and model characteristics in-depth, and implement modifications on data processing and loss function design. These adjustments is verified through extensive evaluations. Nevertheless, we don't leverage information contained within the labels. For future research, how to incorporate the information of individual labels into a multi-label classification model is also a promising direction for us.

References

Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced

bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. *arXiv preprint arXiv:2109.04712*.

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)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Long Ma, Xiaorong Jian, and Xuan Li. 2022. Pai at semeval-2022 task 11: Name entity recognition with contextualized entity representations and robust loss functions. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1665–1670.

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.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.

A Appendix

The team name of us in TIRA system: theodor-zwinger