Arthur Caplan at SemEval-2023 Task 4: Enhancing Human Value Detection through Fine-tuning Pre-trained Models

Xianxian Song, Jinhui Zhao, Ruiqi Cao, Linchi Sui,
Tingyue Guan and **Binyang Li***
Lab of Intelligent Social Computing
University of International Relations, Beijing, China
{kilo0409, jhZhao, rqCao, lcSui,tyGuan, byli}@uir.edu.cn

Abstract

The computational identification of human values is a novel and challenging research that holds the potential to offer valuable insights into the nature of human behavior and cognition. This paper presents the methodology adopted by the Arthur-Caplan research team for the SemEval-2023 Task 4, which entailed the detection of human values behind arguments. The proposed system integrates BERT, ERNIE2.0, RoBERTA and XLNet models with fine tuning. Experimental results show that the macro F1 score of our system achieved 0.512, which overperformed baseline methods by 9.2% on the test set.

1 Introduction

Human values refer to the cognition, understanding, judgment, or choice made based on people's certain thinking and senses, which is a kind of thinking or orientation for people to identify things and determine right and wrong, to reflect a certain value or function of people and objects. Human values are studied both in the social sciences (Schwartz, 1994) and formal argumentation (Bench-Capon, 2003) for decades to observe different value priorities between cultures and disagreement. Some values tend to conflict and others to align, which can cause disagreement on the forward course and opinion.

Several approaches have been developed to study human values in the social sciences. However, the task to identify values in arguments seems challenging due to their large number, often implicit use in arguments, and vague definitions. Thus, this task (Kiesel et al., 2023) is an attempt at the automatic identification of values in written arguments. It classifies a given text parameter and a person's value category as to whether the parameter belongs to the categories. This task uses a set of 20 value categories compiled from the social science literature from their ACL paper (Kiesel et al., 2022).

In this paper, we presented various pre-trained models, such as BERT (Devlin et al., 2018), ERNIE2.0 (Sun et al., 2020), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) to address each component of this task, with a bunch of fine-tuning and ensemble techniques. In particular, we introduced prompt (Liu et al., 2023) in our model. The code of the task has been made publicly available from https://github.com/KiloChips/semeval2023-task4.

In this task, our team finally achieved the following results as shown in Table 1.

Task	F1 score	Rank
main test	0.512	6/41
test-Nahjalbalagha	0.285	11/20
test-nyt	0.322	3/12

Table 1: Outcome and rank.

The rest of the paper is organized as follows: Section 2 introduces the background of the emergence and development of the task. Section 3 presents our system and models. Section 4 describes the experimental setup. Section 5 demonstrates the results and analysis. Finally, we reach the conclusions in Section 6.

2 Background

As for the task of value classification, people have conducted classification studies on values in social sciences for a long time.Rokeach (Rokeach, 1973) showed the correlation between people's ideal state and actual behavior. However, different fields have their own views on the classification of values. rokeach(Rokeach, 1973) analyzed 36 kinds of values from the perspectives of sociology, philosophy, and anthropology. During this period, schwartz(Schwartz, 1994) reflected the multifaceted characteristics of values.

*Corresponding author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1953–1959

schwartz(Schwartz et al., 2012) analyzed 48 kinds of human values from the perspective of the universal needs of individuals and society.

On text recognition of values, these papers (Egan et al., 2016);(Misra et al., 2017);(Chen et al., 2019) studied how to extract author's opinions from essay arguments. After them, Bar (Bar-Haim et al., 2020) and Friedman (Friedman et al., 2021) . proposed a method to get the arguments by analyzing the key points and sentences in the article, thus completing the task of extracting the opinions from the article.

Maheshwari (Maheshwari et al., 2017) represented human personality in a parameterized way, while in these papers (Rahwan et al., 2009); (Teze et al., 2019) analyzed the intrinsic values of a public using the degree of public agreement with a certain point of view. On the basis of the above work, Identifying the Human Values behind Arguments (Kiesel et al., 2022) is the first article identifying the human values behind arguments.This completes the task of calculating and analyzing human values expressed through viewpoints.The classification of values adopts the classification of values in the paper(Schwartz et al., 2012), and the values are divided into 20 subcategories. This paper is the realization and improvement of the evaluation task.

3 System Overview

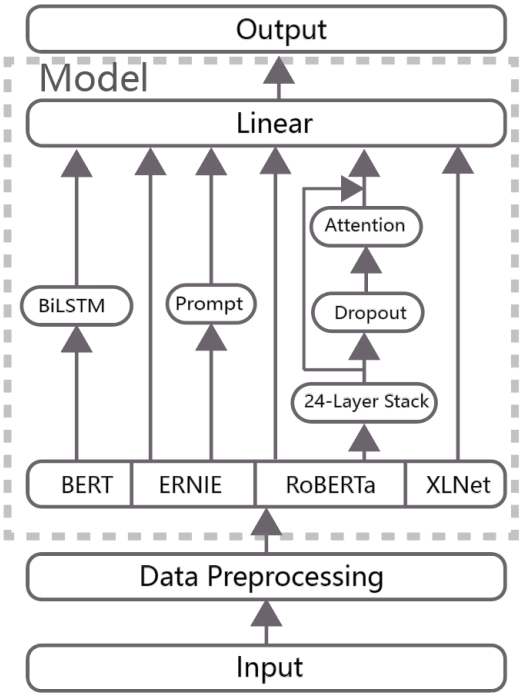

Figure 1: System Architecture.

The presented system architecture, as illustrated in Figure 1,consists of four parts: input, data preprocessing, model, and output. The input section represents the provided raw dataset. The data preprocessing section preprocesses the dataset and inputs the processing results into the model section. In the model part, the input text is represented by vectorization, and then connected to the linear layer for the final multi label classification task. The output part is the classification result predicted by the model, which contains binary labels for 20 different categories.

3.1 Data Preprocessing

The dataset provided by the task organizer consists of two documents: one for arguments and one for corresponding labels. For the convenience of subsequent processing, we merged them into a single document by using **Argument ID** as the common identifier. The merged documents contains attributes such as **Argument ID**, **Conclusion**, **Stance**, **Premise** and **Labels**. The **Labels** refers to the collection of labels for 20 categories to be classified.

To construct the input format for our model, we constructed a new attribute called **Sentence** in the format of "Imagine someone is arguing " **Stance** + "**Conclusion**" + by saying: + "**Premise**". We used **Sentence** as input for our model. Refer to the input section of Table 2 for specific examples.

3.2 Model

We implemented a model ensemble approach, wherein four pre-trained models, namely BERT, ERNIE2.0, RoBERTa, and XLNet, were ensembled.

3.2.1 BERT

BERT-Large (Devlin et al., 2018) is a transformer-based language mode. We used BERT-Large to obtain the pooler embedding of the last hidden layer of the input, which was then fed into a bi-directional LSTM (Zhou et al., 2016) layer. The resulting output was then connected to a fully connected layer for the purpose of classification.

3.2.2 ERNIE2.0

ERNIE2.0 (Sun et al., 2020) is a semantic representation model based on Transformer Encoder. ERNIE2.0 learns real-world semantic knowledge by modeling words, entities, and entity relationships in massive amounts of data. Compared to BERT learning raw language signals, ERNIE2.0

Attribute	content
Conclusion	We should fight for the abolition of nuclear weapons
Stance	against
Premise	nuclear weapons help keep the peace in uncertain times
Sentence	Imagine someone is arguing in favor of “We should fight for the abolition of nuclear weapons” by saying: “Nuclear weapons help keep the peace in uncertain times.”

Table 2: Construction of the attribute of Sentence.

Loss Function	Models
BCE loss	ERNIE2.0+Prompt, RoBERTa+Attention
Softmax+Cross entropy	ERNIE2.0, BERT+BiLSTM, RoBERTa,XLnet

Table 3: The selection of loss function.

directly models prior semantic knowledge units, enhancing the semantic representation ability of the model. We taked the pooler embedding of the last hidden layer in the model and connect it to the fully connected layer for classification.

3.2.3 RoBERTa

RoBERTa (Liu et al., 2019) is an improved version of the BERT model. We used two methods on this model. The first method is the same as what ERNIE2.0 used. In the second method, we first fed the sentence through the pre-trained model and obtained the output of each of the 24 transformer network layers. These outputs were concatenated and weighted according to a set of predefined layer-specific coefficients. Subsequently, we introduced a dropout layer and an attention (Vaswani et al., 2017) layer to calculate the weights of each word. These weights were then multiplied by the previously obtained weighted 24-layer vector representation, yielding a dimension matrix for each word that was subsequently summed. Finally, the resulting values were passed through a linear classification layer.

3.2.4 XLNet

The XLNet model (Yang et al., 2019) cleverly combines the advantages of both LM and BERT models through Permutation Language Modeling. Our experimental approach on this model is the same as ERNIE2.0.

3.3 Prompt

Prompt (Liu et al., 2023) is renowned as the "Fourth Paradigm" in the field of NLP, and is known for facilitating the transfer of pre-trained models to downstream tasks. It eliminates the gap between pre-trained models and downstream tasks by concatenating templates and mapping label words, which has some effectiveness in low-data scenarios.

The template defined in our work is in the form of "*The values implicit in this sentence include:*", which is inserted at the beginning of the original input to form a new input for the model.

3.4 Loss Function

In terms of loss selection, we evaluated the effectiveness of BCE loss and the "softmax + cross entropy" method proposed by Jianlin Su (Su, 2020) for multi-label classification, which is able to alleviate hyperparameter tuning pressure when faced with imbalanced data distribution. Our experimental results demonstrated that both methods exhibited comparable performance on the validation set, with no discernible distinction observed between them. Consequently, we employed both approaches, and the details of their application are documented in Table 5.

3.5 Model Ensemble

Model ensemble is a popular technique in machine learning that involves training multiple models and combining their outputs to improve overall performance. In our study, we employed two methods for model ensemble.

"OR" Operation Due to the substantial class imbalance and a large number of categories, the predicted values of the model tend to be biased towards 0, hence we perform an "*or*" operation on the predicted outputs of the models to be fused. Specifically, if any of the models predict a value of 1, the final prediction is set to 1.

Voting We first categorized the 20 classifications

Type	Category
F1 score > 0.45	Self-direction: action, Achievement, Power: resources, Security: personal, Security: societal, Conformity: rules, Benevolence: caring, Universalism: concern, Universalism: nature
F1 score < 0.45	Self-direction: thought, Stimulation, Hedonism,Power: dominance Face, Tradition, Conformity: interpersonal, Humility, Benevolence: dependability, Universalism: tolerance, Universalism: objectivity

Table 4: The categories based on F1 score.

based on their F1 score on the validation set, with a threshold of 0.45. If the number of models to be fused is greater than 3, we used the following rules: for the higher-scoring class, if two or more models predict a value of 1, the final prediction will be 1. For the lower-scoring classes, we apply the "OR" operation for processing. The specific classification is shown in Table 4. This approach aims to balance the impact of each model and increase the accuracy of the final prediction.

4 Experimental Setup

4.1 Dataset

The SemEval-Task4 dataset[1] is comprised of 20 value categories for which arguments and labels are to be classified. Each argument is composed of three distinct parts, namely promise, instance, and conclusion. The dataset includes a training set consisting of 5,392 arguments, as well as a validation set comprising 1,896 arguments. In addition, three test sets are provided to evaluate the performance of models. The main test set includes 1,576 arguments, while test-Nahjalbalagha contains 279 arguments and is based on Nahjal-Balagha. The argumentation in this dataset varies significantly from the "main" dataset, rendering it more challenging to classify. The specific comparison is shown in the attached table.The test-nyt is a smaller set that includes 80 arguments taken from articles in the New York Times on the subject of coronavirus.

4.2 Parameter Tuning

We undertook fine-tuning of a pre-trained language model, wherein the batch size is set at 64 and the maximum sequence length is limited to 128. The AdamW optimizer is utilized in conjunction with a weight decay factor of 0.01. To account for variations in the pre-training models and methods, varying learning rates and schedulers are employed, as delineated in Table 5.

To facilitate the training process of the RoBERTa+Attention model, we utilized a GPU P100 resource made available by Kaggle[2], whereas the other models are trained on Paddle[3]. The PaddlePaddle-GPU version used for this process must be equal to or greater than 2.4rc, while the PaddleNLP version must be equal to or greater than 2.4.3version.

5 Results

The experimental results of our work are given in Table 6, and the F1 scores for each classification are listed in Table 7. On the main test set, when Bert+BiLSTM, ERNIE2.0, RoBERTa+Attention are ensembled by using OR operation, the F1 score reaches 0.512, which is 9.2% higher than the F1 score of the baseline model. On the New York Times test set, the ensemble of Bert+BiLSTM, ERNIE2.0, RoBERTa can achieve an F1 score of 0.322. 34% is improved over the baseline model's performance. But on the Nahj al-Balagha test set, our method can only achieve an F1 score close to the baseline performance. This shows that there is still room for improvement in the robustness of our method.

From the experimental results, it can be found that a single pre-trained model performs well in classification tasks, and ERNIE2.0 performs best. Therefore, we specifically incorporated the prompt structure into ERNIE2.0. It can be seen that the macro F1 score of the model has dropped a little after adding the prompt. However, in fact, after adding the prompt, the F1 scores of some difficult-to-classify categories (E.g: Face and Universalism: tolerance) have increased. This illustrates the advantage of prompt in low data situations.

[1]https://doi.org/10.5281/zenodo.6814563

[2]https://www.kaggle.com/

[3]https://www.paddlepaddle.org.cn/

Models	learning rate	Epoches	Scheluer	Warmup	Note
BERT+BiLSTM	3e-5	20	LinearDecayWithWarmup	0	
ERNIE2.0	4e-5 2e-5	20 7	\	\	First train 20 epoches with a learning rate of 4e-5, and then train 7 epoches with a learning rate of 2e-5 on the model with the best validation set effect
ERNIE2.0+Prompt	3e-5 3e-4	20	LinearDecayWithWarmup	0	The learning rate of the pre-train model The learning rate of the prompt module
RoBERTa+Attention	3e-5	15	LinearDecayWithWarmup	0	
RoBERTa	4e-5	20	LinearDecayWithWarmup	0	
XLNet	4e-5	20	LinearDecayWithWarmup	0	

Table 5: Parameters tuning of different models.

Test set / Approach	Macro F1 Score
Main	
Baseline	0.422
ERNIE2.0	0.503
*RoBERTa	0.496
*XLNet	0.474
BEPRA-OR	0.506
BEPRA-Voting	0.509
BERA-OR	0.512
*BERAX-Voting	0.517
Nahj al-Balagha	
Baseline	0.279
ERNIE2.0	0.290
BERA-ORBERAX-Voting	0.272
BEPRA-Voting	0.283
BEPRA-OR	0.285
*BERA-OR	0.283
New York Times	
Baseline	0.237
ERNIE2.0	0.320
BEPRA-OR	0.318
BERA-OR	0.322
*BERAX-Voting	0.285

Table 6: The test set result score of the arthur-caplan team. (Baseline is the result of Baseline-BERT provided by the task organizer, which directly uses the BERT model for multi label classification tasks. Data marked with * were not part of the final submission. B is BERT+BiLSTM, E is ERNIE2.0, P is Prompt, R is RoBERTa, A is Attention, X is XLNet. OR and Voting are the strategies we use in our model.)

6 Conclusion

In this paper, we investigated the effectiveness of fine-tuning large pre-trained models for the downstream task of multi-label classification of human values. The results of our experiments indicated that most pre-trained models performed well on this task, with the ERNIE2.0 model exhibiting particularly strong performance. However, we observed that the addition of structures similar to BiLSTM and attention to the pre-trained models often led to a reduction in the models' efficacy. Consequently, our team adopted a model ensemble approach as the primary method for further enhancing the performance of a single pre-trained model in value classification.

7 Acknowledgement

This project was partially supported by National Natural Science Foundation of China (Grant number: 61976066), Beijing Natural Science Foundation (Grant number: 4212031), the Fundamental Research Fund for the Central Universities (Grant numbers: 3262023T19), and Research Funds for NSD Construction, University of International Relations (Grant numbers: 2021GA07).

References

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*.

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

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538*.

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*.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the 3rd Workshop on Argument Mining, The 54th Annual Meeting of the Association for Computational Linguistics*, pages 134–143. Association for Computational Linguistics (ACL).

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. *arXiv preprint arXiv:2110.10577*.

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: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*.

Tushar Maheshwari, Aishwarya N Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. 2017. A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 731–741.

Amita Misra, Brian Ecker, and Marilyn A Walker. 2017. Measuring the similarity of sentential arguments in dialog. *arXiv preprint arXiv:1709.01887*.

Iyad Rahwan, Pavlos Moraitis, and C Reed. 2009. *Argumentation in multi-agent systems*. Springer.

Milton Rokeach. 1973. *The nature of human values.* Free press.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

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.

Jianlin Su. 2020. Extend softmax and multi-label cross entropy to multi-label classification.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Juan CL Teze, Antoni Perello-Moragues, Lluis 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

A Appendix

A.1 More detailed evaluation scores

Table 7 shows the More detailes for each value category as described in Results.

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
ERNIE2.0	.50	.48	.64	.21	.31	.61	.41	.52	.29	.77	.63	.60	.53	.55	.14	.55	.26	.72	.79	.41	.47
*ERNIE2.0 Prompt	.49	.49	.61	.20	.28	.58	.40	.48	.31	.75	.62	.54	.48	.49	.13	.53	.26	.74	.79	.42	.44
BEPRA-Voting	.51	.48	.63	.27	.30	.56	.44	.45	.30	.77	.63	.57	.49	.51	.16	.54	.34	.73	.81	.40	.54
BEPRA-OR	.51	.51	.64	.25	.30	.62	.43	.52	.30	.73	.61	.59	.55	.47	.16	.54	.35	.71	.79	.40	.55
BERA-OR	.51	.53	.65	.26	.30	.62	.43	.52	.29	.73	.62	.61	.56	.48	.16	.54	.34	.72	.80	.40	.54
Nahj al-Balagha																					
Best per category	.63	.62	.60	.50	.67	.68	.48	.56	.62	.73	.73	.55	.63	.34	.47	.66	.58	.74	.67	.54	.59
Best approach	.57	.62	.60	.43	.27	.68	.48	.56	.44	.73	.73	.44	.63	.34	.47	.66	.58	.74	.57	.54	.59
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
BERAX-Voting	.27	.07	.29	.13	.22	.60	.10	.00	.49	.39	.20	.52	.36	.15	.20	.38	.18	.27	.00	.05	.25
BEPRA-OR	.28	.10	.23	.14	.22	.61	.12	.19	.52	.48	.17	.53	.22	.19	.24	.24	.20	.22	.29	.08	.28
BEPRA-Voting	.28	.12	.26	.00	.27	.61	.13	.00	.55	.51	.20	.55	.24	.14	.23	.25	.19	.21	.33	.10	.29
New York Times																					
Best per category	.50	.50	.22	.00	.03	.54	.40	.00	.50	.59	.52	.22	.33	1.00	.57	.33	.40	.62	1.00	.03	.46
Best approach	.34	.22	.22	.00	.00	.48	.40	.00	.00	.53	.44	.00	.18	1.00	.20	.12	.29	.55	.33	.00	.36
BERT	.24	.00	.00	.00	.00	.29	.00	.00	.00	.53	.43	.00	.00	.00	.57	.26	.27	.36	.50	.00	.32
1-Baseline	.15	.05	.03	.00	.03	.28	.03	.00	.05	.51	.20	.00	.07	.03	.12	.12	.26	.24	.03	.03	.33
ERNIE2.0	.32	.14	.15	.00	.00	.31	.00	.00	.50	.59	.30	.00	.18	1.00	.15	.11	.36	.40	.29	.00	.42
BERA-OR	.32	.13	.09	.00	.00	.46	.17	.00	.40	.56	.26	.00	.15	1.00	.14	.13	.33	.40	.25	.00	.36
BEPRA-OR	.32	.14	.10	.00	.00	.46	.17	.00	.40	.56	.27	.00	.18	1.00	.14	.14	.34	.36	.29	.00	.38

Table 7: Achieved F_1-score of team arthur-caplan per 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.