Philo of Alexandria at Touché: A Cascade Model Approach to Human Value Detection

Notebook for the Touché Lab at CLEF 2024

Víctor Yeste[1,2,*], Mariona Coll-Ardanuy[1] and Paolo Rosso[1,3]

[1]PRHLT Research Center, Universitat Politècnica de València, 46022, Valencia, Spain
[2]Universidad Europea de Valencia, 46010, Valencia, Spain
[3]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)

Abstract

This paper describes our contribution to the Human Value Detection shared task at CLEF 2024. Our submitted system approaches the task of human value detection and attainment using a sequence of two models: a multi-label text classifier based on DeBERTa is used first to predict the human values present in the text. Then, a follow-up natural language inference binary classifier based on DeBERTa is applied to discern whether the values that are present in the text are attained or constrained. This cascade model approach improves the granularity of text classification. Our approach outperforms all baselines, achieving a Macro F1-score of 0.28 on sub-task 1 (human value detection) and a Macro F1-score of 0.82 on sub-task 2 (value attainment prediction).

Keywords

human value detection, text classification, multi-label classification

1. Introduction

The task of human value detection involves applying natural language processing to identify whether human values are present in texts, and to determine whether such values appear as attained or constrained. These values have been ordered in a circular motivational continuum by Schwartz et al. (2012) [1], in which 19 values were defined based on their compatible and conflicting motivations, expression of self-protection vs. growth, and personal vs. social focus.

The Human Value Detection at CLEF 2024 task (*ValueEval'24*) [2] consists of two sub-tasks: the first is to detect the presence or absence of each of these 19 values, while the second is to detect whether the value is attained or constrained. The dataset provided for both tasks consist of approximately 3000 human-annotated texts between 400 and 800 words created by the ValuesML project [3]. The data is provided at the sentence-level (44,758 sentences for training, 14,904 sentences for validation, and 14,569 sentences are kept for testing), in which each sentence is annotated in a multi-label setting and a single-level taxonomy consisting of 38 labels, expressing each human value's attained and constrained versions. As the original dataset is multilingual and contains texts in several languages, an automatically translated version to English of the training, validation, and test dataset was provided for every team that wished to create an approach without a multilingual perspective.

The present work includes a cascade model approach consisting of two consecutive models: a multi-label text classifier used to predict which of the 19 human values are present in the text, followed by a binary classifier which treats the task of determining the attainment or not of the value as a stance classification problem, in which both the text and the value are passed as input, and the expected output is whether the value appears as attained or constrained. Our approach outperforms all the baselines provided by the organizers, including a baseline based on BERT. This paper includes a detailed system overview, the experiments we have performed, the results and discussion, and some conclusions and

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
vicyesmo@upv.es (V. Yeste); mcoll@prhlt.upv.es (M. Coll-Ardanuy); prosso@dsic.upv.es (P. Rosso)
https://victoryeste.com (V. Yeste)
0000-0002-3660-8347 (V. Yeste); 0000-0001-8455-7196 (M. Coll-Ardanuy); 0000-0002-8922-1242 (P. Rosso)

future studies that could continue this work. The code for the proposed system, as well as for all our experiments, is available on GitHub.[1]

2. System Overview

This section presents our cascade model approach, in which two models are dedicated to each of the proposed sub-tasks, and combined to achieve the prediction in the required format. Our approach uses the automatic translated texts to English. Our system introduces a cascade model approach for the detection and stance classification of the predefined set of human values. It consists of two subsystems: one for detecting the presence of each human value and another for establishing the stance (if the sentence attains or constrains) of each human value. Each subsystem is fine-tuned separately, in both cases using a DeBERTa model[2] [4] as base, for the task of sequence classification using the HuggingFace implementation.[3]

- **Subsystem 1:** Its primary function is to identify the presence of human values within sentences. By combining the 'attained' and 'constrained' labels to indicate an overall presence, it streamlines the multi-label classification task, simplifying it to a binary classification for each of the 19 human values (presence vs. absence). The model for the proposed subsystem is available at HuggingFace.[4]
- **Subsystem 2:** it receives the outputs of subsystem 1 and classifies the stance towards each present human value in a binary classification (attained vs. constrained). This system transforms the sentences dataset into premise-hypothesis pairs, where each sentence is the *premise*, a value is the *hypothesis*, and the 'attained' and 'constrained' labels are the *stance*. The model for the proposed subsystem is available at HuggingFace.[5]

Given that subsystem 1 focuses on detecting the presence of human values in the text, and subsystem 2 focuses on the stances towards each detected human value, this cascade model approach improves the granularity of text classification. As can be seen in the Results section, it also enhances the performance of the final predictions.

3. Experiments

Experiments were carried out on Google Colab in Python 3.10.12 and Nvidia Tesla, as well as 12.7 GB of System RAM and 15 GB of GPU RAM. HuggingFace transformers [5] have been used as frameworks for all the experiments in this study. Training has been designed with flexibility and performance, and evaluation metrics have been calculated upon training completion and validation with the task validation dataset. F1 scores for each label and a macro-average F1 score were used to evaluate each experiment, enabling a comprehensive analysis of individual and overall effectiveness.

3.1. Preliminary Experiments

Our initial experiments involved using a single model approach to classify each text into the predefined set of human value stance labels (i.e., the 38 labels determining whether the sentence attains or constrains each of the 19 human values). The objective was to leverage the powerful features of well-known transformer models for this purpose, and to determine which was the best suited for the task. We experimented with the following pre-trained models: `google-bert/bert-base-uncased` [6],[6] `FacebookAI/roberta-base`[7]

[1] https://github.com/VictorMYeste/touche-human-value-detection
[2] https://huggingface.co/microsoft/deberta-base
[3] https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModelForSequenceClassification
[4] https://huggingface.co/VictorYeste/deberta-based-human-value-detection
[5] https://huggingface.co/VictorYeste/deberta-based-human-value-stance-detection
[6] https://huggingface.co/google-bert/bert-base-uncased
[7] https://huggingface.co/FacebookAI/roberta-base

[7], `microsoft/deberta-base`[8] [4], `google/electra-base-discriminator`[9] [8] and `xlnet-base-cased`[10] [9]. These pre-trained models were initialized for sequence classification and, for task 1, configured for the multi-label classification setting.

Each selected model was fine-tuned on the task training dataset and validated with the task validation dataset. The sentences were tokenized using the specific tokenizer from Huggingface Transformers for each model. All models were fine-tuned with a batch size of 8, for 5 training epochs, a learning rate of $2e{-}5$, and a weight decay 0.01. A linear learning rate scheduler was implemented using 0 warmup steps on BERT and RoBERTa, using Adam as an optimizer and incorporating weight decay directly to improve regularization and prevent overfitting. The final model, DeBERTa, was selected based on the fact that it produced the highest macro F1-score on the training and validation dataset.

3.2. System Experiments

Our cascade models approach has been developed by fine-tuning two DeBERTa models in sequence, therefore converting the approach of dividing the challenge into two sub-tasks into reality. In both cases, we used the same experimental settings as described in the preliminary experiments section.

First, we transformed each pair of attained and constrained labels into presence labels, understanding presence as an OR operation between both labels. The DeBERTa model was fine-tuned for multi-label classification of the 19 available human values, and was trained on the task training dataset and validated on the task validation dataset to evaluate the effectiveness of this subsystem alone to detect the presence of human values. This step ensures the ability to answer the first sub-task of the challenge with a significantly reduced complexity as the output space is 19-dimensional instead of 38-dimensional, translated into a smaller number of possible label combinations.

Second, subsystem 2 receives subsystem 1 results as inputs and applies an approach of natural language inference, where each sentence is considered a premise, human values labels are considered different hypotheses, and "attained" and "constrained" are the labels. With this technique, the model tries to determine a logical entailment relationship between this pair of sequences. This inference establishes the stance of the sentence toward each human value, which answers sub-task 2 of the proposed challenge.

Finally, it is important to note that, in order to adjust the predictions of our cascade approach to the format required by the shared task, we had to do one small modification to our system. While our system is conceived to apply the second model only for those values that have been found to be present in the text, the format required to participate in both tasks[11] meant that, in order to produce our results file, we applied the subsystem 2 model to each sentence-value pair, instead of only those values that have been predicted to be in the sentence. To ensure that values detected as absent remain below the 0.5 threshold that is used by the evaluator to determine that the value is not present, in those cases in which the value has not been predicted by the first model, we multiply the second model prediction score by the first model prediction score, divided by two.

4. Results

In our preliminary experiments, our models were trained and evaluated with the provided training and validation datasets, generating an individual F1-score for every human label and a generic Macro F1-score, which were used to compare the effectiveness of the different models. The model with

[8] https://huggingface.co/microsoft/deberta-base

[9] https://huggingface.co/google/electra-base-discriminator

[10] https://huggingface.co/xlnet/xlnet-base-cased

[11] Only one file had to be submitted for both tasks, with 38 columns for each of the 38 labels (i.e. 19 human value pairs). Task 1 was evaluated based on the sum of values between the attained and constrained columns of the value (which should be larger than 0.5 if the value is present), and task 2 was evaluated based on which of the two columns ('attained' or 'constrained') had the larger value. The organizers recommended avoiding setting the same number for both attained and constrained, even if our system predicted that the value was not referenced in the text.

Table 1
Achieved F_1-score (0.score) of each submission on the test dataset for subtask 1. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

		F_1-score																		
Submission	**EN**	**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**
philo-of-alexandria (our approach)	✓	28	08	22	27	31	35	31	34	17	33	40	47	42	09	00	21	28	40	57	21
valueeval24-bert-baseline-en	✓	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02
valueeval24-random-baseline		06	02	07	05	02	11	08	10	04	05	13	03	11	03	00	04	04	09	04	02
valueeval24-random-baseline	✓	06	02	07	05	02	11	08	10	03	04	14	03	11	03	00	05	04	09	04	02

the highest effectiveness was found to be DeBERTa with a Macro F1-Score of 0.20. However, while DeBERTa presented the highest Macro F1-score, some models achieved higher individual F1-scores for some human values: BERT was better on 'tradition attained'; RoBERTa on 'achievement attained', 'security: societal constrained', 'universalism: concern attained', and 'universalism: nature attained'; Electra on 'power: dominance attained', 'power: resources constrained', 'security: societal attained', 'universalism: concern attained', and 'universalism: concern constrained'; and XLNet on 'power: resources attained', 'power: resources constrained', 'security: societal attained', 'conformity: rules constrained', 'benevolence: dependability attained', 'universalism: concern attained', and 'universalism: concern constrained'. These results could indicate that using a different model for each human value could be an interesting approach. As DeBERTa was selected as the best overall model, our system was developed using two cascade DeBERTa models.

Table 1 shows the results of our system for subtask 1. As it can be seen, our system outperforms all baselines, including the BERT-based baseline, by 0.04 in terms of F1-score. It is interesting to note that both our approach and the BERT baseline generally perform similarly well on the same values (such as 'security: societal', 'tradition', 'conformity: rules', and 'universalism: nature'), and similarly bad on other values (such as 'self-direction: thought', 'conformity: interpersonal', and 'humility'), while some other values have significant increases with our approach (such as 'universalism: tolerance' and 'face'). Overall, our approach matches or outperforms the BERT baseline for all values, except for 'power: resources'.

Table 2 shows the results of our system for subtask 2. While our approach outperforms the BERT baseline, the F1-score is only slightly higher (0.82 over 0.81). Our approach only outperforms the BERT baseline on 12 of the 19 possible values. Our model is best at predicting 'hedonism' and 'benevolence: caring', and significantly worse than the baseline in predicting 'humility', with which our first model also failed.

5. Conclusions

This work proposes a system to resolve the challenge sub-tasks related to human values detection. Our approach uses cascade DeBERTa models, where the first detects the presence of each human value, and the second detects if the sentence attains or constrains the present human values in each sentence. The latter approach improves the effectiveness of the baseline at the test dataset by 4 on sub-task 1 and by 1 on sub-task 2. These models were trained on a subset of 44,758 sentences in English, validated on a subset of 14,904 sentences, and tested on a separate subset of 14,569 sentences.

Table 2
Achieved F_1-score (0.score) of each submission on the test dataset for subtask 2. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

		F_1-score																		
Submission	**EN**	**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**
philo-of-alexandria (our approach)	✓	82	85	80	85	91	86	79	80	78	85	80	82	77	78	77	93	89	84	83	79
valueeval24-bert-baseline-en	✓	81	83	79	86	88	84	77	80	74	84	81	78	78	79	87	89	86	85	81	78
valueeval24-random-baseline		53	55	49	52	54	52	56	56	50	48	54	50	54	55	61	55	51	48	51	51
valueeval24-random-baseline	✓	52	51	47	54	52	53	55	53	52	52	50	54	53	49	45	53	56	52	49	56

Future work could involve implementing a separated detection model for each human value, adapting each model to its characteristics depending on which model performs better in each case. Considering the complexity and subtlety of this task, adding linguistic and statistical characteristics to texts could enrich their context and improve the effectiveness of the models.

Acknowledgments

Work for this paper was conducted as part of the PhD Program in Computer Science at the Universitat Politècnica de València. The work of Mariona Coll Ardanuy and Paolo Rosso was funded by the research project FairTransNLP, grant PID2021-124361OB-C31, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

[1] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, et al., Refining the theory of basic individual values, Journal of personality and social psychology 103 (2012) 663.

[2] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[3] The ValuesML Team, Touché24-ValueEval, 2024. doi:10.5281/zenodo.10663363.

[4] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2020.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing: system demonstrations, 2020, pp. 38–45.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers

for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[8] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2019.

[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).