Edward Said at Touché: Human Value Detection Using Transformers and Upsampling

Notebook for the Touché Lab at CLEF 2024

Aisha Nur Aydin, Shaden Shaar and Claire Cardie

[1]*Cornell University*

Abstract

In this paper, we tackle both subtasks of the proposed shared task Human Value Classification at Touché– that aims to classify dialogue speech into one of 19 human values determined by Schwartz's Refined Theory of Basic Individual Values. We fine-tune models like DeBERTa and RoBERTa with F1-loss to handle multi-label settings. We additionally test different sampling strategies to accommodate for data imbalance. We found that by training on the English-translated utterances, we beat the baselines by at least 2 F1 points across both subtasks.

Keywords

human-value-classification, Touché, CLEF

1. Introduction

The Human Value Detection task aims to identify the values humans express through words. Eight languages are included in the task, so the human value is meant to be identified in a multilingual context. Upon identification of the values, the value can be classified as having been attained, constrained, or neither attained nor constrained. The first subtask pertains to classifying a sentence to contain a specific human value. In contrast, the second subtask identifies whether the sentence constrains or attains the given set of human values[1].

We fine-tuned pre-trained RoBERTa and DeBERTa models based on the data from the ValuesML dataset.[1] We observed that the class imbalance in the data could affect the classification of the human values, so we attempted various combinations of upsampling the data to address this issue. We found that upsampling the lowest-performing categories by 4 folds yields the best results. Similar to other tasks (e.g., emotion detection [2]), we use only the English models and train using English and translated English utterances. Using the same model trained for both tasks, we were able to beat all the proposed baselines.

2. Background

There are nineteen human values in Schwartz's Refined Theory of Basic Individual Values [3]. There is a wide range of values in this taxonomy. Additionally, since human values can be addressed implicitly, it can be challenging to identify which value is being used in a given text [4]. The human value detection task for 2023 had three components that made up each argument: a conclusion, a stance, and a premise. Here is an example argument from the 2023 Human Values Task:

Conclusion	We should ban human cloning
Stance	in favor of
Premise	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.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

ana72@cornell.edu (A. N. Aydin); ss2753@cornell.edu (S. Shaar); ctc9@cornell.edu (C. Cardie)

[1]ValuesML project

This year, the task takes a sentence as an input and outputs its human value, whether it is (even partially) attained, constrained, or neither. A value is attained if the text supports it and is constrained if it hinders it. This required us to treat the problem as a multi-class, multi-label classification problem with 38 labels consisting of 19 human values that are attained or constrained.

An example sentence with a human value is the following. The input is:

> *"Young women examining their options for third level education have been urged to consider careers in science, technology, engineering or maths (STEM)."*

The value referred to in this sentence (the output) is Achievement Attained.

These sentences are provided by the ValueML dataset , which contains text and their respective human values from articles and political text. The data consists of 3000 texts in over eight languages. 20% of the data is used for validation, 20% is part of testing, and the remaining 60% is for training.

3. System Overview

For all our experiments, we chose to combine all languages in the dataset into one by keeping the English utterances only. We use RoBERTa Large, specifically "FacebookAI/roberta-large" as the base model for our submission. We first experimented with the original RoBERTa and DeBERTa models but found the F1 scores on the validation dataset lower than the BERT baseline provided by the task organizers. We then fine-tuned RoBERTa-Large and DeBERTa-Large and observed improved results. There was an improvement, but the overall F1 score appeared to be affected by the class imbalance. Human Values that appeared less in the dataset had noticeably lower F1 scores than those that seemed more.

We attempted multiple configurations of upsampling but ended up using an upsampling technique where the human values with lower performance metrics were chosen to be upsampled by a factor of four.

To determine which human values to upsample, we fine-tuned the RoBERTa large model on the training data and analyzed the evaluation results. If the F1 score was 0.15 or less for subtask 1, we upsampled the attains and constrains form of that human value. We also looked at the metrics for subtask 2 to determine whether we should upsample only one of the attains or constrains of that value. If the recall for one of the constrains or attains was 50% or less than its counterpart, we only increase the underperforming one. For example, if the recall of the attains version of a human value was 50% or less than the constrains version, we only choose to upsample the attains version and vice versa. Based on the metrics and the unevenness of the dataset, we upsampled these values by a factor of four:

- Self-direction: thought constrained
- Self-direction: action constrained
- Humility attained
- Humility constrained
- Face attained
- Face constrained
- Benevolence: caring constrained
- Benevolence: dependability constrained
- Universalism: tolerance attained
- Universalism: tolerance constrained
- Conformity: interpersonal attained
- Conformity: interpersonal constrained
- Tradition constrained
- Power: dominance constrained

Figure 1 illustrates the impact of upsampling on the dataset, showing the distribution of data before and after upsampling.

4. Experimental Setup

To fine-tune the model, we set the learning rate to 2e-5, used a warm-up ratio of 0.2, set the batch size to 8, and used four epochs. We put the random seed to 42, used the AdamW optimizer, and used a

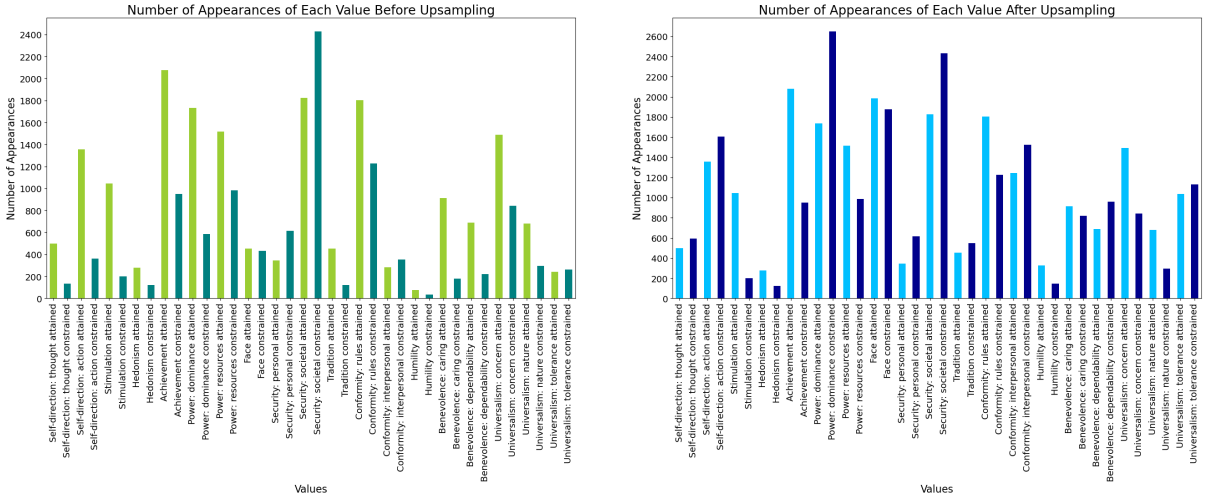

Figure 1: Comparison of data distribution before and after upsampling. Left: Before Upsampling, Right: After Upsampling.

linear scheduler. We used one A100 GPU to run the experiments for the final submitted models. Our experiments used pre-trained RoBERTa [5] and DeBERTa [6] models. For evaluation, we use Precision, Recall, and the macro F1-score. We fine-tuned four models: RoBERTa and DeBERTa large on the upsampled data, and RoBERTa and DeBERTa large on the regular data, with no upsampling. For our final submission, we chose the model fine-tuned with the upsampled data using RoBERTa large as the base. This had the best metrics among all four models on the validation dataset. All four models can be found on Hugging Face[2].

Table 1
Achieved F_1-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**
muted-glacier-2024-05-07-02-06-56 (RoBERTa large with Upsampling)	✓	28	05	17	11	15	25	31	34	16	32	41	45	44	06	05	10	23	41	57	27
other-models-2024-07-05-04-47-24 (DeBERTa large with Upsampling)	✓	26	03	14	26	18	30	12	20	21	16	43	52	46	06	07	09	13	36	58	19
other-models-2024-07-05-04-47-48 (DeBERTa large no Upsampling)	✓	26	00	17	07	09	38	29	27	19	24	44	48	45	00	00	18	11	42	58	11
other-models-2024-07-05-04-48-29 (RoBERTa large no Upsampling)	✓	23	00	12	04	26	27	26	18	07	18	41	39	44	00	00	16	04	39	57	06
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

5. Results

The upsampling methods resulted in a 4% increase for the macro F1 score of subtask 1 and a 2% increase for the macro F1 score of subtask 2. Of the upsampled human values, "Benevolence: caring" and

[2]https://huggingface.co/collections/aishanur/human-value-detection-668c4548607e863cc5cebd58

Table 2
Achieved F_1-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**
muted-glacier-2024-05-07-02-06-56 (RoBERTa large with Upsampling)	✓	83	77	82	85	88	88	79	80	77	84	84	85	80	80	76	90	86	85	85	78
other-models-2024-07-05-04-47-24 (DeBERTa large with Upsampling)	✓	84	81	80	84	92	89	77	82	78	85	86	89	81	76	68	92	89	86	84	79
other-models-2024-07-05-04-47-48 (DeBERTa large no Upsampling)	✓	85	81	83	85	90	90	81	82	76	86	85	84	81	85	87	93	88	86	83	86
other-models-2024-07-05-04-48-29 (RoBERTa large no Upsampling)	✓	84	81	83	86	93	89	80	80	78	86	84	84	81	87	84	90	89	85	82	84
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

"Tradition" were the only ones that did not have an improved F1-score for subtask 1. For the other upsampled values, there were improved F1-scores, some marginal and others significant. For subtask 2, the upsampled values performed only marginally better except for "Self-direction: thought" and "Humility", which all had worse metrics, and "Benevolence: dependability" and "Universalism: tolerance", which stayed the same.

After the task deadline, we looked at the performances of the other three models on the test data. Those models were DeBERTa large fine-tuned on upsampled data and DeBERTa large and RoBERTa large fine-tuned on the original training data. The results comparing all of these models are on Table 1 and Table 2. The RoBERTa large model with upsampling performs better than the other three models in subtask 1, but ends up being the worst-performing model on subtask 2. The best performing model for subtask 2 is DeBERTa large without upsampling.

6. Conclusion and Future Work

We were able to beat the BERT baselines by incorporating the F1-Loss function and up-sampling lower-performing categories. This entails that the data benefits from knowledge transfer obtained from the different categories. In the future, we would also like to consider different languages as just using the translation could have led to missed cultural nuances expressed in language.

References

[1] 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.

[2] S. Hassan, S. Shaar, K. Darwish, Cross-lingual emotion detection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo,

J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 6948–6958. URL: https://aclanthology.org/2022.lrec-1.751.

[3] 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). doi:10.1037/a0029393.

[4] J. Kiesel, M. Alshomary, N. Mirzakhmedova, M. Heinrich, N. Handke, H. Wachsmuth, B. Stein, SemEval-2023 Task 4: ValueEval: Identification of Human Values behind Arguments, in: R. Kumar, A. K. Ojha, A. S. Doğruöz, G. D. S. Martino, H. T. Madabushi (Eds.), 17th International Workshop on Semantic Evaluation (SemEval 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2287–2303. doi:10.18653/v1/2023.semeval-1.313.

[5] 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.

[6] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention (2020). arXiv:2006.03654.