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] European University of Valencia, 46010, Valencia, Spain

[3] Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)

SYSTEM OVERVIEW

- A **cascade model approach for the detection and stance classification of the predefined set of human values**
- Two subsystems are dedicated to each of the proposed sub-tasks combined to achieve the prediction in the required format. Each subsystem is fine-tuned separately, in both cases using a **DeBERTa[1] model as base**, for the task of sequence classification using the HuggingFace implementation.
- They use the subset of automatically translated texts into **English.**

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

SUBSYSTEM 1

- Its primary function is to **identify the presence of human values within sentences**.
- Combining the 'attained' and 'constrained' labels to indicate an overall presence, simplifying the multi-label classification task to a binary classification for each of the **19 human values (presence vs. absence).**
- The model for the proposed subsystem is available at HuggingFace[2].

[2] https://huggingface.co/VictorYeste/deberta-based-human-value-detection

SUBSYSTEM 1

Submission	EN	F_1-score																			
		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

SUBSYSTEM 2

- It receives the Subsystem 1 outputs 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[3].

[3] https://huggingface.co/VictorYeste/deberta-based-human-value-stance-detection

SUBSYSTEM 2

Submission	EN	F_1-score																			
		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

SCORING SUBTASK 2

- Our system is conceived to apply the second model only to those values present in the text.
- The format required to participate in both tasks meant that to produce our results file, **we applied the subsystem 2 model to each sentence-value pair** instead of only those values predicted to be in the sentence.
- To ensure that values detected as absent remain below the 0.5 threshold that the evaluator uses to determine that the value is not present, we take a specific approach: In those **cases in which the first model has not predicted the presence of the value**, we multiply the second model prediction score by the first model prediction score, divided by two.

RESULTS

- **The model with the highest effectiveness was DeBERTa** with a Macro F1-Score of 0.20. However, **other models achieved higher individual F1-scores for some human values**.
- Our **system for subtask 1** outperforms all baselines, including the BERT-based baseline, by 0.04 in terms of F1-score (0.28). Our approach matches or outperforms the BERT baseline for all values except for 'power: resources.'
- Our **system for subtask 2** outperforms the BERT baseline, but the F1-score is only slightly higher (0.82 over 0.81). It outperforms the BERT baseline on 12 of the 19 possible values.

CONCLUSIONS

- Future work could involve **implementing a separate detection model for each human value** and adapting each model to its characteristics, depending on which model performs better.
- 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.

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)