Beyond Single Models: Leveraging LLM Ensembles for Human Value Detection in Text

Diego Dimer Rodrigues, Mariana Recamonde-Mendoza, Viviane P. Moreira

[1]Instituto de Informática – (UFRGS)

{ddrodrigues,mrmendoza,viviane}@inf.ufrgs.br

Abstract. *Every text may reflect its writer's opinions, and these opinions, especially in political contexts, are often tied to specific human values that they either attain or constrain. Identifying these values can provide policymakers with deeper insights into the underlying factors that influence public discourse and decision-making. While current large language models (LLMs) have shown promise across various tasks, no single model may generalize sufficiently to excel in tasks like human value detection. In this work, we utilize data from the Human Value Detection task at CLEF 2024 and propose leveraging multiple ensembles of LLMs to enhance the identification of human values in text. Our results found that the ensemble models achieved higher F1 scores than all baseline models, suggesting that combining multiple models can offer performance comparable to very large models but at much lower memory requirements.*

1. Introduction

People can agree or disagree on numerous topics even when using the same information. These differences arise largely from their individual beliefs about what is worth striving for, a concept referred to as (human) *values*. Human values can conflict or align, leading to a wide range of opinions on controversial issues. This divergence is one of the reasons for the formation of different political parties, each representing the values of specific groups [Kiesel et al. 2022].

Given its significance, the study of human values spans multiple disciplines, including social sciences [Schwartz 1994] and formal argumentation [Bench-Capon 2003]. Researchers focused on various aspects, such as classifying values, detecting them in text, and understanding their societal impact. In computer science, there is a growing body of work dedicated to value detection and emotion recognition from text [Dellaert et al. 1996, Tariq et al. 2019, Ammanabrolu et al. 2022]. These tasks are challenging and yet have a broad spectrum of applications, such as aiding policymakers in gauging public sentiment, detecting political alignment, and more.

In this work, we aim to advance the field of human value detection by leveraging multiple ensembles of Large Language Models (LLMs) to identify these values in text and enhance model performance. We adopt the value taxonomy presented in [Schwartz et al. 2012], which categorizes values into two types for each value—attained and constrained. However, our task focuses solely on identifying the presence of a value in a sentence, so we sum the attained and constrained versions to determine whether a sentence contains a particular value. We conduct this study with a dataset from CLEF 2024. The data is highly imbalanced, making this a challenging classification problem.

2. Background and Related Work

Human Value Detection has recently gained attention, particularly as the focus of a shared task at CLEF 2024. This task aimed to detect human values in speech, attracting participation from 20 teams. The outcomes of this competition, including the performance metrics of each team, are detailed in [Kiesel et al. 2022]. These efforts underscore the complexity of detecting nuanced human values in text and highlight the need for advanced models that can accurately capture such subtleties.

LLMs have revolutionized NLP tasks across various domains. The introduction of Transformer architectures [Vaswani et al. 2017] marked a significant leap forward, leading to the development of powerful pre-trained models like BERT [Devlin et al. 2019], RoBERTa [Liu et al. 2019], and DeBERTa [He et al. 2021]. These models have been highly effective in text classification, sentiment analysis, and content generation, significantly reducing the need for training models from scratch. Numerous studies [Xian et al. 2023, Hoang et al. 2019, Sun et al. 2019, Sobhanam and Prakash 2023] have demonstrated the efficacy of fine-tuning these models for specific tasks, showcasing their versatility and robustness in handling diverse NLP challenges.

Ensemble Learning is a well-established technique in machine learning, often employed to improve predictive performance by combining multiple models. Traditionally associated with decision trees [Quinlan 1986], ensemble learning has evolved to incorporate various frameworks, including those involving LLMs [Jiang et al. 2023].

3. Methodology

The data used in this study comes from the Human Value Detection at CLEF (Conference and Labs of the Evaluation Forum) 2024 task (ValueEval'24) [Kiesel et al. 2024a] and consists of approximately 3K human-annotated texts containing over 73K sentences. The annotation associated with each sentence indicates whether a specific human value is "attained" and "constrained". A total of 19 human values are analyzed. Each column receives the value 0, 0.5, or 1, indicating whether the sentence does not contain the human value, partially contains it, or fully contains it, respectively. This study focused on the English dataset. All models were optimized for F1-Macro score.

To approach the task as a multi-label classification problem, we combined the "attained" and "constrained" columns in the *labels* file, summing their values to determine whether a specific human value is present in a sentence (0 for false, 1 for true). The result was an array of 19 boolean values for each sentence, which were then used as inputs for model fine-tuning. Thus, each human value represents a class and the predictive model may assign more than one class for a given sentence. While the value *Humility* was removed by many CLEF participants due to its scarcity in the training set (present in only 0.2% of sentences), we retained it, considering it important to predict even rare values to ensure comprehensive performance across all values.

Using the training dataset, we fine-tuned six models: base and large versions of BERT [Devlin et al. 2019], RoBERTa [Liu et al. 2019], and DeBERTa [He et al. 2021]. After fine-tuning, we used the validation data to create a new dataset that included the sentences, prediction probabilities for each class, and binary predictions indicating whether a value is present in a sentence. The true labels are also carried onto the dataset to enable evaluations. Five different ensemble approaches were used to combine model outputs:

- **prob-equal**: Probabilities from each model were summed and then averaged. A threshold of 0.2 was applied.
- **prob-large-double**: Probabilities from base models were summed, and probabilities from large models were doubled before summing. The total was divided by the number of votes (nine), and a threshold of 0.2 was applied.
- **preds-majority**: Binary predictions from all models were summed, with a threshold of 2 applied to predict a value as present if at least two models identified it.
- **preds-large-double**: Binary predictions were summed, with large models receiving two votes each. A threshold of 2 was used, meaning a value would be predicted as present if one large model or two base models identified it.
- **prob-weight-macro-f1**: The probabilities predicted by each model were weighted by their F1 scores on the validation set. The weighted probabilities were then summed and normalized, followed by applying a threshold of 0.2.

For reproducibility, all experiments, ensemble diagrams, and scripts used for fine-tuning are available on GitHub[1], with a fixed random seed for all libraries. Implementation details and further results are also in our repository. The models used in this study are publicly accessible and can be downloaded from HuggingFace.

4. Results

Results are presented in Table 1. The RoBERTa Large model achieved the highest accuracy among the individual models, which aligns with expectations given the larger model size. However, since the primary metric for model selection during training was the macro F1-score rather than accuracy, it is not surprising that larger models and ensemble models do not consistently show higher accuracy.

Table 1. F1 and Accuracy results for our models and baselines. ⋆ means the model is an ensemble, and † means it used the multilingual dataset version

	Model	Macro F1	Accuracy
Base models	BERT-base-uncased	0.160	0.502
	BERT-large	0.263	0.482
	RoBERTa-base	0.248	0.485
	RoBERTa-large	0.282	0.508
	DeBERTa-base	0.274	0.480
	DeBERTa-large	0.295	0.507
Ensembles	prob-equal	0.330	0.447
	prob-large-double	0.326	0.438
	prob-weight-macro-f1	0.330	0.445
	preds-majority	0.318	0.484
	preds-large-double	0.319	0.418
Baselines	[Legkas et al. 2024] †	0.390	–
	[Yunis 2024] ⋆ †	0.350	–
	[Yeste et al. 2024]	0.280	–

Table 1 also compares our results with the top-3 models from the CLEF 2024 submissions. Notably, our ensemble approaches, specifically *prob-weight-macro-f1* and

[1] https://github.com/diegodimer/valueeval24

prob-equal, performed only 0.03 and 0.02 below the top-scoring models from the conference, which utilized XLM models and the multilingual dataset. The approach by Arthur Schopenhauer [Yunis 2024] leveraged an ensemble of DeBERTa-v2-xxlarge and xlmRoBERTa-large models. Similarly, Hierocles of Alexandria [Legkas et al. 2024] employed both the multilingual and English-translated datasets, incorporating sentence sequence information and fine-tuning an XLM-RoBERTa-xl model. Finally, team Philo of Alexandria [Yeste et al. 2024] fine-tuned a DeBERTa model specifically for this task.

Looking into the scores for each of the 19 human values, we see that our ensembles demonstrated competitive performance, closely matching the results of XLM models and outperforming the DeBERTa-base model across nearly all values. This task was particularly challenging due to the significant class imbalance in the dataset, with nearly 50% of test set instances not containing any of the 19 values. This imbalance skews predictions towards false negatives, resulting in lower F1 scores despite high accuracy, as models may correctly predict the absence of values due to their prevalence.

Overall, the results demonstrate that ensemble models can achieve performance comparable to very large models, even when utilizing models that require less computational resources. Although training an XLM-DeBERTa model was not feasible on the hardware used for this study due to memory constraints, our ensembles still achieved a strong macro F1-score. Specifically, the best ensemble model improved the macro F1-score from 0.295 (the highest among the base models) to 0.33, highlighting the effectiveness of ensemble methods in enhancing model performance in this context.

5. Conclusion

In this study, we tackled the complex task of identifying human values in text, a challenge crucial for understanding the values that shape public discourse and decision-making. By leveraging multiple ensembles of LLMs, we demonstrated that ensemble-based approaches could significantly enhance individual model performance in this task. This suggests that instead of relying solely on a single, powerful LLM, ensemble methods offer a more robust and effective solution for complex NLP tasks.

Despite the advanced capabilities of models like GPT-4.0, these models still struggle to consistently deliver satisfactory performance in this domain. For instance, in the ValueEval'24, a team using GPT-4.0 for zero-shot classification achieved an F1-score of 0.25 [Kiesel et al. 2024b], which is lower than the performance of our ensemble approaches. This highlights the inherent challenges in human value detection, where the nuances of language and context often exceed the capacity of a single model, no matter how sophisticated. Future work will include a qualitative analysis to better understand the errors made by the models and improve the proposed approaches, reinforcing the potential of ensemble learning as a key strategy in advancing the field.

Acknowledgments. This work has been partially funded by CNPq-Brazil and Capes Finance Code 001.

References

[Ammanabrolu et al. 2022] Ammanabrolu, P., Jiang, L., Sap, M., Hajishirzi, H., and Choi, Y. (2022). Aligning to social norms and values in interactive narratives. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

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

[Dellaert et al. 1996] Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1970–1973 vol.3.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.

[He et al. 2021] He, P., Liu, X., Gao, J., and Chen, W. (2021). DEBERTA: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

[Hoang et al. 2019] Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-based sentiment analysis using BERT. In Hartmann, M. and Plank, B., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.

[Jiang et al. 2023] Jiang, D., Ren, X., and Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion.

[Kiesel et al. 2022] Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., and Stein, B. (2022). Identifying the Human Values behind Arguments. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

[Kiesel et al. 2024a] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., De Longueville, B., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzhakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., and Stein, B. (2024a). Overview of touché 2024: Argumentation systems. In Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., and Ounis, I., editors, *Advances in Information Retrieval*, pages 466–473, Cham. Springer Nature Switzerland.

[Kiesel et al. 2024b] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., Longueville, B. D., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., and Stein, B. (2024b). Overview of Touché 2024: Argumentation Systems. In Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Nunzio, G. M. D., Soulier, L., Galuscakova, P., Herrera, A. G. S., Faggioli, G., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

[Legkas et al. 2024] Legkas, S., Christodoulou, C., Zidianakis, M., Koutrintzes, D., Petasis, G., and Dagioglou, M. (2024). Hierocles of alexandria at touché: Multi-task & multi-head custom architecture with transformer-based models for human value detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS. org*.

[Liu et al. 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

[Quinlan 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[Schwartz 1994] Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45.

[Schwartz et al. 2012] Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., and Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

[Sobhanam and Prakash 2023] Sobhanam, H. and Prakash, J. (2023). Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification. *International Journal of Information Technology*, 15(7):3669–3677.

[Sun et al. 2019] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

[Tariq et al. 2019] Tariq, Z., Shah, S. K., and Lee, Y. (2019). Speech emotion detection using iot based deep learning for health care. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4191–4196.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Xian et al. 2023] Xian, G., Guo, Q., Zhao, Z., Luo, Y., and Mei, H. (2023). Short text classification model based on DeBERTa-DPCNN. In *2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 56–59.

[Yeste et al. 2024] Yeste, V., Ardanuy, M., and Rosso, P. (2024). Philo of alexandria at touché: A cascade model approach to human value detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS. org*.

[Yunis 2024] Yunis, H. (2024). Arthur schopenhauer at touché 2024: Multi-lingual text classification using ensembles of large language models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS. org*.