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

anonymous authors

[1]address

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 through machine learning 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 demonstrate 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 with significantly lower memory requirements.*

1. Introduction

People can disagree or agree on numerous topics even when using the same information to form opinions. These differences arise largely from their individual beliefs about what is worth striving for and how to achieve it, a concept referred to as (human) *values*. Human values can sometimes 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 in these fields have 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 application, 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[1]. 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

[1]Attained means that whatever is described in the sentence will help lead to fulfilling the value. In contrast, an event can be stated in a way that hinders (or constrains) the value.

2024[2], which includes manually annotated texts by over 70 individuals. The data is highly imbalanced across the various values, making this a challenging classification problem.

2. Theoretical Background

In this work, we utilize three pre-trained language models – BERT, RoBERTa, and DeBERTa – each in their base and large versions (the latter having twice the number of Transformer layers). These models are well-suited for different natural language processing (NLP) tasks: BERT is robust and versatile for general NLP applications, RoBERTa excels when extensive pre-training data and computational power are available, and DeBERTa is particularly effective in tasks involving complex linguistic structures and nuanced contextual information.

The Transformer architecture, introduced by [Vaswani et al. 2017], underpins many modern NLP models, including those used in this study. A key innovation of Transformers is their ability to process sequences in parallel, a significant advancement over previous models that handled sequences sequentially. This parallelism has led to substantial improvements in the performance of various NLP tasks, making Transformers a foundational element in developing models like BERT, RoBERTa, and DeBERTa.

The goal of this work is to leverage these models with their strengths and particularities and demonstrate that with both base and large versions, which require relatively fewer computational resources, one can build ensemble models that achieve performance levels comparable to more resource-intensive models like XLM. For example, our models were trained on GPUs with under 12GB of VRAM, whereas loading an XLM model with a larger batch size might not even be feasible within such constraints. This section provides a brief overview of the three LLMs employed in our ensemble models.

BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al. 2019], is a pre-trained model that generates bidirectional encoder representations by considering the context from both the left and right sides of a target word in all attention layers. This bidirectional approach allows BERT to effectively understand word context within sentences. The BERT model used in this work was pre-trained on tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

RoBERTa (A Robustly Optimized BERT Pre-training Approach) [Liu et al. 2019], builds upon the BERT architecture with several key enhancements to improve performance. While retaining BERT's bidirectional nature, RoBERTa is trained on larger datasets using bigger mini-batches and more robust optimization techniques. Additionally, RoBERTa removes the NSP task, which was found not to contribute to the model's performance, and introduces dynamic masking of tokens during training.

DeBERTa (Decoding-enhanced BERT with Disentangled Attention) [He et al. 2021], further refines the Transformer architecture by addressing some of the limitations of both BERT and RoBERTa. DeBERTa introduces two key innovations: a disentangled attention mechanism, where embeddings are represented by two vectors (content and position), and an enhanced masked decoder, which allows the model to

[2] https://touche.webis.de/clef24/touche24-web/human-value-detection.html

capture richer contextual information. These improvements make DeBERTa particularly effective in tasks that require a deeper understanding of word context and structure, leading to superior performance across various NLP tasks.

3. Related Work

This section provides an overview of the key areas relevant to our study: human value detection, LLMs, and ensemble learning.

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 the work by [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. For instance, in [Jiang et al. 2023], an ensemble approach is used to address the variability in LLM performance across different examples, demonstrating that a pair ranker ensemble can outperform individual LLMs and baseline methods. Similarly, [Fang et al. 2024] presents a state-of-the-art LLM ensemble model for Product Attribute Value Extraction in E-commerce, where the outputs of individual models are aggregated to produce superior predictions, outperforming even the best single models on Walmart's internal dataset. Additionally, [Abburi et al. 2023] explores the application of well-known LLMs such as BERT, DeBERTa, and RoBERTa in generating probabilities that feed into traditional machine learning models, thereby enhancing the final prediction accuracy.

4. Methodology

This section outlines the data used in our study and describes the six individual models and five ensemble models evaluated in our experiments.

4.1. Data

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. 2024] 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

Table 1. Hyper-parameters used in fine-tuning. The symbol '*' in the model version means it was used for both versions (large and base)

Hyper-parameter	Model Version	Value
Training Epochs	base	15
	large	10
Tokenizer max length	base	128
	large	256
Learning Rate	*	1e-5
Weight decay	*	0.01
Threshold	*	0.5
Optimizer	*	Adam with learning rate 1e-5

"attained" and "constrained". For example, considering the human value *Self-direction: thought*, two columns related to it are included, *Self-direction: thought attained* and *Self-direction: thought 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.

The dataset is pre-split into training, validation, and test sets by the ValueEval team. It is available in two formats: a multilingual version with sentences in multiple languages and an English version where all data is translated into English. For our study, we fine-tuned the models using the English dataset. The training set was used for model training, the validation set for metric computation, and the final metrics, discussed in Section 5, were calculated on the test set. All models were optimized for the F1-Macro score, as this metric was chosen by the task organizers for ranking participant approaches.

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.

During data processing, we identified and removed duplicated texts from both the training and validation sets. Additionally, sentences with fewer than 10 characters were excluded to minimize noise. The data was then tokenized with padding to the maximum length, as described in Table 1, using the tokenizer specific to each pre-trained model. 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, and assuming that its inclusion could impact the performance of the other values.

4.2. Proposed Approach

To combine the outputs of different LLMs, we fine-tuned six models on the training dataset, running 10 epochs for the large versions and 15 for the base versions. The hyper-parameters used in the fine-tuning process are detailed in Table 1. After fine-tuning, we created from the validation data a new dataset that included the sentences, prediction probabilities (the direct output from the models, representing the probability that a sen-

tence belongs to each class), and binary predictions (0 or 1, indicating whether a value is present in a sentence) from each LLM. The true labels are also carried onto the dataset to evaluate the predictions. The structure of this dataset is depicted in Figure 1.

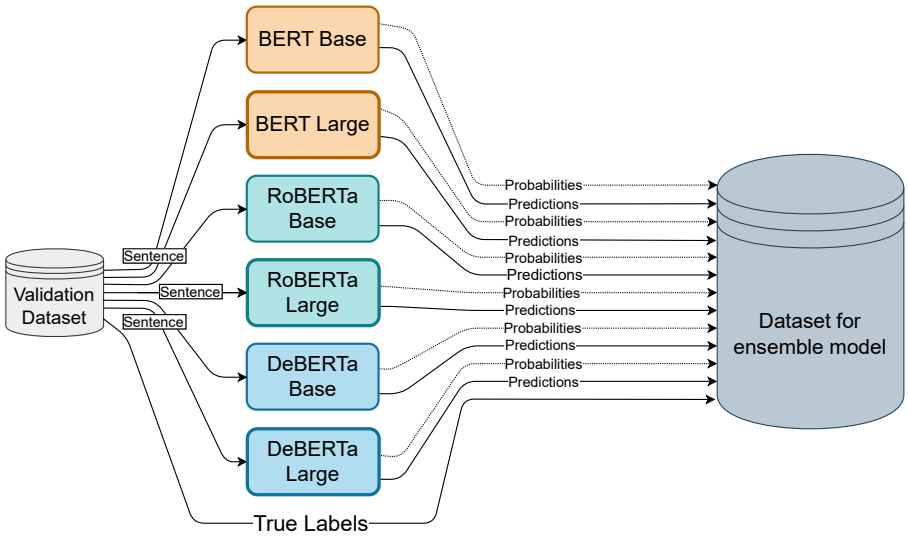


Figure 1. Base ensemble of fine-tuned LLMs

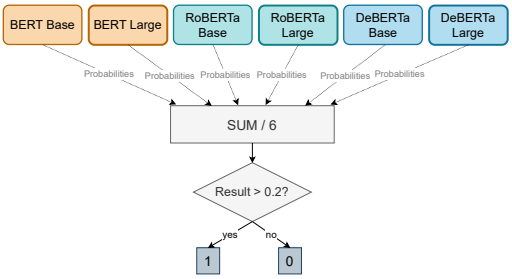
To combine the output of the models, this work proposes five different approaches, as depicted in Figure 2: three using the probabilities and two using the final predictions. Threshold values were empirically decided while testing different possible values.

- **prob-equal**: Probabilities from each model were summed and then averaged by dividing by the number of models (six). A threshold of 0.2 was applied to determine whether a value is present in the sentence (see Figure 2(a)).
- **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 to determine value presence (see Figure 2(b)).
- **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 model identified it (see Figure 2(d)).
- **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 (see Figure 2(c)).
- **prob-weight-macro-f1**: The probabilities predicted by each model were weighted by its F1 score on the validation set. The weighted probabilities were then summed and normalized, followed by applying a threshold of 0.2 to decide value presence (see Figure 2(e)).

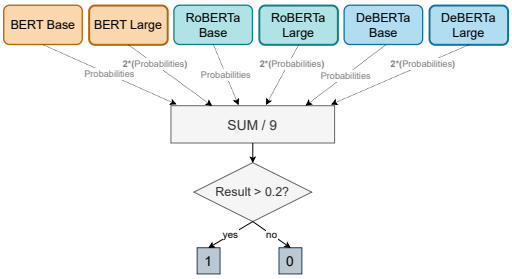
For reproducibility, all experiments, ensemble diagrams, and scripts used for fine-tuning are available on GitHub[3], with a fixed random seed for all libraries. The models used in this study are publicly accessible and can be downloaded from HuggingFace[4].

[3] suppressed to preserve anonymity

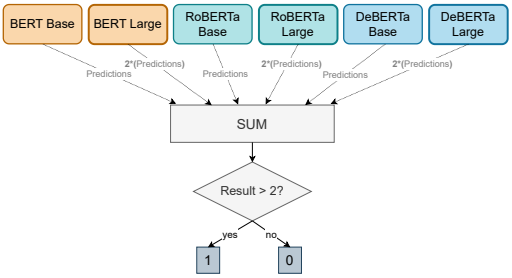
[4] `https://huggingface.co/`



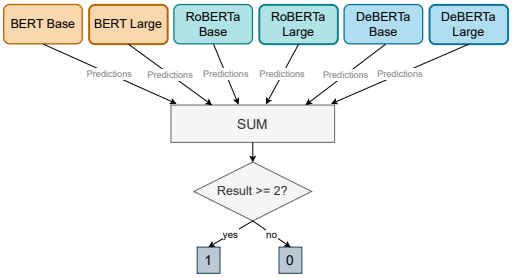
(a) Prob-equal approach



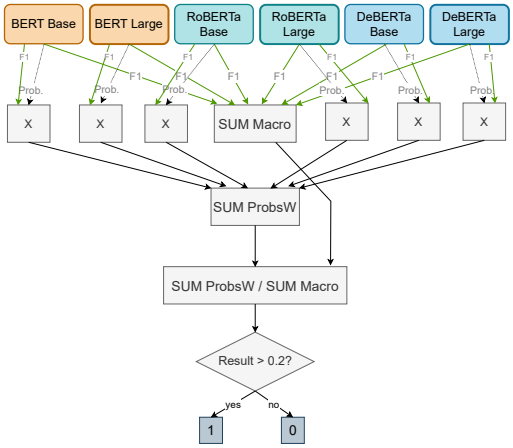
(b) Prob-large-double approach



(c) Preds-large-double approach



(d) Preds-majority approach



(e) Prob-weight-macro-f1 approach

Figure 2. Ensemble-based approaches adopted in the current work.

5. Results

Evaluation results in terms of accuracy and F1-score are presented in Table 2. 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 2 also compares our results with the top-3 models from the CLEF 2024 submissions. Notably, our ensemble approaches, specifically *prob-weight-macro-f1* and *prob-equal*, performed only 0.03 and 0.02 below the top-scoring models from the con-

Table 2. 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	–

ference, which utilized XLM models and the multilingual dataset. The approach by Arthur Schopenhauer [Yunis 2024][5] leveraged an ensemble of DeBERTa-v2-xxlarge and xlmRoBERTa-large models. Similarly, Hierocles of Alexandria [Legkas et al. 2024][6] 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][7] fine-tuned a DeBERTa model specifically for this task.

Table 3 provides the macro F1-scores for each of the 19 human values predicted in this task. Our models demonstrated competitive performance, closely matching the results of XLM models and outperforming the DeBERTa-base model across nearly all values. The value "Humility" was omitted from CLEF by the participant's choice since this value had very little representation in the training and validation sets, leading to a macro F1-score of 0 for those models. 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 model predictions towards false negatives, which results in lower F1-scores despite potentially high accuracy, as models may correctly predict the absence of values simply 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] https://github.com/h-uns/clef2024-human-value-detection
[6] https://github.com/SotirisLegkas/Touche-ValueEval24-Hierocles-of-Alexandria
[7] https://github.com/VictorMYeste/touche-human-value-detection

Table 3. F1-Scores for each human value

	[Legkas et al. 2024]	[Yunis 2024]	[Yeste et al. 2024]	prob-weight-macro-f1	preds-large-double
Self-direction: thought	0.15	0.12	0.08	0.11	0.11
Self-direction: action	0.27	0.24	0.22	0.23	0.22
Stimulation	0.30	0.33	0.27	0.30	0.28
Hedonism	0.37	0.35	0.31	0.31	0.32
Achievement	0.45	0.40	0.35	0.39	0.39
Power: dominance	0.42	0.37	0.31	0.35	0.34
Power: resources	0.49	0.47	0.34	0.39	0.37
Face	0.31	0.24	0.17	0.29	0.28
Security: personal	0.42	0.38	0.33	0.38	0.36
Security: societal	0.49	0.46	0.40	0.44	0.43
Tradition	0.46	0.49	0.47	0.49	0.44
Conformity: rules	0.51	0.50	0.42	0.45	0.44
Conformity: interpersonal	0.24	0.19	0.09	0.16	0.15
Humility	0	0	0	0.05	0.04
Benevolence: caring	0.34	0.32	0.21	0.28	0.25
Benevolence: dependability	0.33	0.31	0.28	0.32	0.33
Universalism: concern	0.47	0.46	0.40	0.42	0.40
Universalism: nature	0.63	0.60	0.57	0.59	0.58
Universalism: tolerance	0.27	0.27	0.21	0.24	0.22

6. 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[8], 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.

[8]https://ceur-ws.org/Vol-3740/paper-322.pdf

References

[Abburi et al. 2023] Abburi, H., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E., and Bhattacharya, S. (2023). Generative AI text classification using ensemble LLM approaches.

[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.

[Fang et al. 2024] Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., and Achan, K. (2024). Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction.

[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. 2024] 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. (2024). 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.

[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*.