

Hierocles of Alexandria at Touché: Multi-task & Multi-head Custom Architecture with Transformer-based Models for Human Value Detection

Notebook for the Touché Lab at CLEF 2024

Sotirios Legkas^{1,†}, Christina Christodoulou^{1,†}, Matthaios Zidianakis^{1,†},
Dimitrios Koutrintzes^{1,†}, Maria Dagioglou^{1,†} and Georgios Petasis^{1,†}

¹Institute of Informatics & Telecommunications, National Centre for Scientific Research (N.C.S.R.) ‘Demokritos’, Aghia Paraskevi, Attica, Greece

Abstract

The paper presents our participation as *Hierocles of Alexandria* in Touché at CLEF 2024, which addressed the *Human Value Detection* shared task. The objectives of the task was to detect one or more human values (sub-task 1) and their attainment (sub-task 2) in lengthy texts across nine languages, including the automatic translation of these texts into English. Our methodology involved the fine-tuning of four Transformer language models within a customized multi-head model architecture for multi-label text classification. The experimental approach comprised comprehensive data analysis, the utilization of various loss functions, and class positive weights to handle class imbalance. Additionally, we incorporated previous sentences as context and represented human values as special tokens in the texts to enhance classification performance. Notably, all our submissions for the multi-lingual data surpassed the baseline submissions in both sub-tasks 1 and 2. Our top-performing submission secured the 1st position among all the participating teams in sub-task 1 in both the multi-lingual and English-translated data.

Keywords

human values, multi-label text classification, custom multi-head architecture, multi-lingual, transformers

1. Introduction

Values motivate our actions [1] and impact all processes of our (moral) behaviour from perception and judgment to focus and action [2]. Being essentially the driving forces of individuals and societies, intelligibly identifying them empowers us to understand more profoundly, among others, our cultural heritage [3], citizen’s political behaviour [4] and human interaction with artificial agents [5]. This knowledge can be fed back to people through the delivery of sustainable and responsible solutions from the related duty holders. Naturally, narratives are vessels of values. Historical texts, social media content, news items, and ChatGPT products are all resources to extract values and inform research, resolve sociopolitical tensions, deliver responsible AI. In particular, Touché [6] aims to advance our understanding of decision-making and opinion-forming processes by supporting the development of related methods and tools based on human values detection.

Human values detection in natural language is a complex task due to diverse perceptions, multilingualism, terminology interpretation, values attainment and actor attribution, among others. These are challenges that we have encountered through our research [7, 8] and our participation in relevant projects^{1,2}, and are also reflected in the performance of the models developed as part of SemEval-2023

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

[†] Authors contributed equally

 sotirislegkas@iit.demokritos.gr (S. Legkas); ch.christodoulou@iit.demokritos.gr (C. Christodoulou); mzidianakis@iit.demokritos.gr (M. Zidianakis); dkoutrintzes@iit.demokritos.gr (D. Koutrintzes); mdagiogl@iit.demokritos.gr (M. Dagioglou); petasis@iit.demokritos.gr (G. Petasis)

 0009-0000-9468-5650 (S. Legkas); 0009-0009-5616-3937 (C. Christodoulou); 0009-0009-5102-8103 (M. Zidianakis); 0009-0003-7401-6347 (D. Koutrintzes); 0000-0002-3357-2844 (M. Dagioglou); 0000-0003-3157-1597 (G. Petasis)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.vast-project.eu/>

²https://knowledge4policy.ec.europa.eu/projects-activities/valuesml-unravelling-expressed-values-media-informed-policy-making_en

Task 4: ValueEval [9]. Touché [6] at CLEF 2024 provides the opportunity to examine many of the aforementioned challenges. The provided dataset is a collection of 3000 human-annotated texts, including news articles and political texts, chosen to reflect diverse views. Over 70 people, from 9 language teams, annotated texts (in their mother tongue) for their value content and attainment. Schwartz's values [1] were adopted for the annotation vocabulary. In addition to the original files provided in nine languages, a machine-translated version of them in English was provided.

The current state of the art in Natural Language Processing (NLP) and ML/AI has enabled the development of methods that identify human values in natural language artifacts [10, 11, 12, 13]. Advanced techniques for text classification, particularly for shorter text sequences, rely on fine-tuning Transformer-based models [7], Large Language Models (LLMs) [14], ensembles of Transformer-based models [15, 16], and custom model architectures involving multiple heads with attention mechanisms [17, 18].

This paper presents the methodology and results of Hierocles of Alexandria team in the *Human Value Detection* shared task of the second edition of Touché [6] at CLEF. Motivated by the availability of longer texts in this Touché edition and the availability of multilingual data, the innovative aspects of our approach include the modeling of contextual information, and the application of multi-task learning. Assuming that the value classification of a sentence may depend on earlier sentences and their classifications, previous sentences and their labels (either from annotated data during training or from classification of previous sentences during evaluation) are provided as input along with the sentence under classification. Multi-task learning, in the form of language-specific classification tasks, has been employed in order to capture potential different value instantiations in different languages.

Our approach leveraged fine-tuning four Transformer-based language models within a custom multi-task with multiple heads model architecture, specifically tailored for multi-lingual and multi-label text classification in order to capture the linguistic nuances. Our experimental strategy comprised a comprehensive data analysis, the application of various loss functions, and the utilization of class positive weights to mitigate the challenge of class imbalance. Our approach achieved the highest score, securing the 1st place for both multilingual and English submissions in sub-task 1, surpassing all other participating teams and baselines. The code for our approach is available on the provided GitHub link.³

In the context of Touché, the results from all approaches were submitted through the TIRA platform, which ensured the reproducibility and reliability of the software employed by participants, thereby facilitating the comparison of information retrieval experiments [19].

The structure of this paper is as follows: Section 2 explores the background. Several aspects of the data, including data analysis, pre-processing and an exploratory phase, are presented in Section 3. Section 4 introduces an overview of the developed system and the experiments. Section 5 presents the results. Finally, in Section 6 the conclusions are discussed, including limitations and future work.

2. Background

The exploration of human values within Natural Language Processing (NLP) encompasses various theoretical and empirical endeavors. Central to this exploration is Shalom H. Schwartz's theory of basic human values, which identifies nineteen universal values inherent to human behavior and cultural expression. These values, driven by distinct motivational goals, form a circular structure that illustrates their dynamic interplay, where pursuing one value may align with or conflict with another [1]. Schwartz's framework provides valuable insights into the motivational goals driving human actions and the complex interrelations among different values. In NLP, this framework offers a robust foundation for identifying and interpreting human values embedded within language.

In a significant research endeavor focused on identifying human values in NLP, a comprehensive taxonomy comprising 54 human values was crafted, aligning closely with psychological research. The researchers also introduced the initial annotated dataset for studying human values behind arguments [20]. This dataset encompassed 5.270 arguments from four distinct cultures: Africa, China, India, and the

³<https://github.com/SotirisLegkas/Touche-ValueEval24-Hierocles-of-Alexandria>

USA. Each argument in the dataset consisted of a premise, a conclusion, and a stance attribute indicating whether the premise supported or opposed the conclusion. The researchers manually annotated these arguments for human values. Their methodology has paved the way for automating the classification of human values, yielding promising results, with F1-scores reaching up to 0,81 and averaging 0,25, establishing a benchmark for future research in this domain.

To further advance the field of human values detection in argumentative texts, the authors of the aforementioned research organized the *ValueEval: Identification of Human Values Behind Arguments* shared task 4 in SemEval-2023 [9] by mapping the 54 human values from their previous research to a set of 20 value categories for multi-label classification. The task showcased both the potential and challenges associated with identifying human values in argumentative texts. A total of 39 teams contributed their methodologies, utilizing the Touché23-ValueEval Dataset comprising 9.324 arguments sourced from 6 diverse outlets, including religious texts, political forums, free-text arguments, newspaper editorials, and online democracy platforms in English [21]. Each argument included a premise, a conclusion, and a stance attribute signifying whether the premise was in favor of or against the conclusion. The teams' approaches were primarily evaluated on the Macro-F1 score.

The task's winner, the Adam-Smith team, achieved an F1 score of 0,56 by calculating a global decision threshold during training that optimizes the F1 score. They mainly employed twelve individual Transformer-based models that are ensembled in order to perform multi-label classification. [16]. The second-place John-Arthur team found that it is beneficial to encode the input data by adding tokenizer's special token separators, corresponding to low-cardinality values of Stance (in favour of vs against). Also, they fine-tuned larger Language Models, which performed better. Lastly, they adopted a threshold of 0,2 at the output of the sigmoid function to get the binary predictions for each human value, achieving an F1 score of 0,55 [14]. Addressing implicit value discrimination and data imbalance, the PAI team employed a multi-label classification model with a class-balanced loss function, securing multiple top positions across task categories with an overall average score of 0,54, placing them third [15]. The Mao-Zedong team's introduction of a multi-head attention mechanism and a contrastive learning-enhanced K-nearest neighbor mechanism resulted in an F1 score of 0,53, placing them fourth [17]. Finally, certain members of the Hierocles of Alexandria team, who participated in that year's task as part of the Andronicus of Rhodes team, leveraged a Transformer model with four classification heads and applied two classification strategies with different activation and loss functions. In addition, they used two different data partitioning methods to handle class imbalance. Their system, employing majority voting, achieved an F1 score of 0,48, placing them in the upper half of the competition [7].

Inspired by the best methodologies employed in ValueEval, our approach aimed to tackle class imbalance and improve the classification performance, through the use of sigmoid threshold, larger language models, and tokenizer's special tokens in the encoded input. Nevertheless, Touché's Human Value Detection task at CLEF 2024 [6] has extended human value detection by integrating multiple languages, besides English. The introduction of new languages introduces more challenges, such as possible differences in annotation styles among languages, which adds complexity to the problem. To this end, our proposed approach was customised to the dataset features and the task by incorporating techniques that address the issue of multi-linguality and capture the linguistic nuances. The introduction of multiple languages and the need to address language-specific phenomena, was the main motivation behind our proposed approach in this paper, which includes a model architecture with multiple heads that are specifically tailored to each language, aiming to model more accurately multi-label text classification across multiple languages.

3. Data

3.1. Data Analysis

The dataset comprises 2.648 complete texts in nine languages: English, Greek, German, French, Bulgarian, Hebrew, Italian, Dutch, and Turkish. The dataset is split by the shared task organisers into training, validation and test sets: Of these texts, 1.603 are used for training, 523 for validation, and 522 for testing.

The number of annotated texts per language varies, as illustrated in Table 1. English has the highest number of texts (408), while French and Hebrew have the fewest (219 and 250, respectively). Each text is segmented into sentences, resulting in 74.231 sentences: 44.758 for training, 14.904 for validation, and 14.569 for testing.

The number of labels varies for each language, as shown in Table 1. There is no correlation between the number of texts and the number of labels. For instance, even though Hebrew texts were among the fewest, they had the highest number of labels (4.992).

For sub-task 1, the texts are annotated with Schwartz's 19 personal values [1]. Nearly half of the sentences (30.662 out of 59.662) are labelled with one or more values. Labels for the test set are not provided to evaluate participating systems. For sub-task 2, each classified value includes an annotation indicating whether the value is attained or constrained by the sentence, resulting in a final dataset with 38 classes.

The frequency of human value labels varies quite a bit, as depicted in Figure 1. *Security: societal* is the most frequently used value, with over 5.000 labels. *Achievement* and *Conformity: rules* are also quite popular, with over 3.000 labels each. On the other hand, *Self-direction: thought*, *Universalism: tolerance*, and *Humility* are less commonly used, with fewer than 1.000 labels. In fact, *Humility* is the least represented value, with only 151 labels.

All data was provided in the original language and translated using the DeepL API, except for Hebrew, which was translated using the Google Translate API.

Table 1
Number of texts and labels per language.

	All	English	Greek	German	French	Italian	Dutch	Bulgarian	Turkish	Hebrew
No. Texts	2648	408	328	261	219	276	323	260	323	250
No. Labels	27038	2590	3658	3876	1111	3262	3265	3624	4284	4992

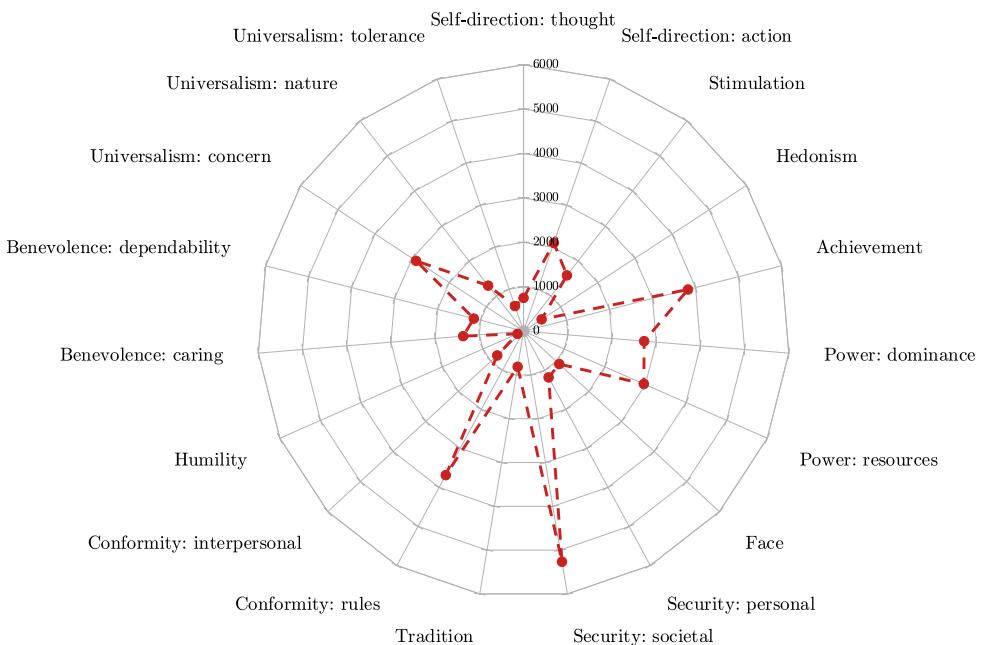


Figure 1: Labels' frequency across all texts. Some labels like "Security societal", "Conformity: Interpersonal" and "Achievement" are more frequent within the dataset.

3.2. Data Pre-processing

The dataset provides unique identifiers for each sentence, indicating the source text and its position within that text. Despite individual sentences being labeled independently, annotators were presented with the complete text for annotation, potentially influencing their assessments based on the overarching context.

To address this consideration, our pre-processing approach focuses on adding contextual information by integrating the previous part of the text and its annotated values. This was achieved through two main strategies: incorporating previous sentences and adding special tokens.

1. *Incorporating Previous Sentences.* For each sentence, we appended the two preceding sentences to each target sentence, thereby providing context from the specific text. If the total number of tokens exceeded the maximum allowed by our base model (maximum: 512), tokens were removed starting from the most distant sentence. If the sentence was the first sentence of a text, no preceding sentences were added. The “</s>” separator token linked the preceding and target sentences.
2. *Adding Special Tokens.* We implemented special tokens to represent each class, such as “<Security: societal>” representing the value “Security: societal”. We used the annotated labels from the previous two sentences for each sentence and appended them to the end. No special token was added if there was no annotated label for the previous sentences. This enables the classifier to interpret the annotator’s perspective for better contextual understanding. These tokens were added as special tokens in the model’s tokenizer and the token embedding matrix of the model was resized. They were assigned to attributes in the tokenizer for easy access and to make sure they were not split during tokenization. For predictions on the validation and test sets, the predicted classes were used as special tokens to enhance the model’s contextual understanding.

The following is an example of pre-processed English text for the model input:

[CLS] Having spoken to many different left-leaning Hispanics, Avila said, “they are really beginning to feel like the Democratic party has become too extreme to the point where it’s starting to scare some of them.” <Security: societal> </s> Many are beginning to turn away from the Democratic party because “they’re getting vibes of a communist Cuba and socialist Venezuela here in America.” As a result, Avila said Hispanics are going to be “extremely instrumental” in the upcoming midterm elections. <Self-direction: action> </s> “They are starting to come to the realization that their conservative values are in opposition to what the media has been trying to feed them in favor of Biden and the Democrats.” [SEP]

3.3. Exploratory Phase

We carried out several experiments to explore the behaviour of the pre-trained language models in order to form a baseline for our development process. This phase primarily exploits the language models to assess how well the collective human values of the mentioned dataset are captured by pre-trained Transformer models [18], enhanced with classification heads so as to perform multi-label text classification, given the respective textual inputs. This process ensures that the models are fine-tuned to adequately fit the dataset with respect to all the language and sentence constraints.

The baseline experiments involved both multi-lingual (all languages together) and mono-lingual (each language separately) tests in order to record the effect of the special traits of each language on the human values. In general, we observed that the multi-lingual performance of the baseline models on the human value classification is higher than the performance on the individual languages in the mono-lingual experiments, as shown in Table 4 of Appendix A. This outcome could be explained by the close relation of the several inherent features (e.g context, vocabulary) of each language to the human value perception.

To further inspect this correlation and to note the bias of each language, we developed a more specific architecture that is built upon the baseline models, in order to improve the performance.

4. System Overview

4.1. Model Architecture

Based on the findings in Section 3.1, Figure 1 illustrates the imbalance in label distribution across different languages. This imbalance is partly due to the varying annotation styles among languages. For instance, some languages, such as English, have a large number of examples but fewer annotations, while others, like Hebrew, have fewer examples but many annotations. Consequently, similar sentences in different languages may receive different annotations. Experimental results indicate that a single out-of-the-box pre-trained Transformer model fails to effectively capture the unique linguistic features of each language, given the different annotation styles per language. In contrast, models that are fine-tuned for a specific language outperform those that are fine-tuned across all languages.

To address the multi-lingual nature of the problem and the differences in annotations between each language, a custom ensemble model was constructed. The architecture, as seen in Figure 2, leverages a pre-trained Transformer language model as its foundation. On top of this, nine custom Transformer heads were added, each tailored to a specific language: English, Greek, Dutch, Turkish, French, Bulgarian, Hebrew, Italian, and German.

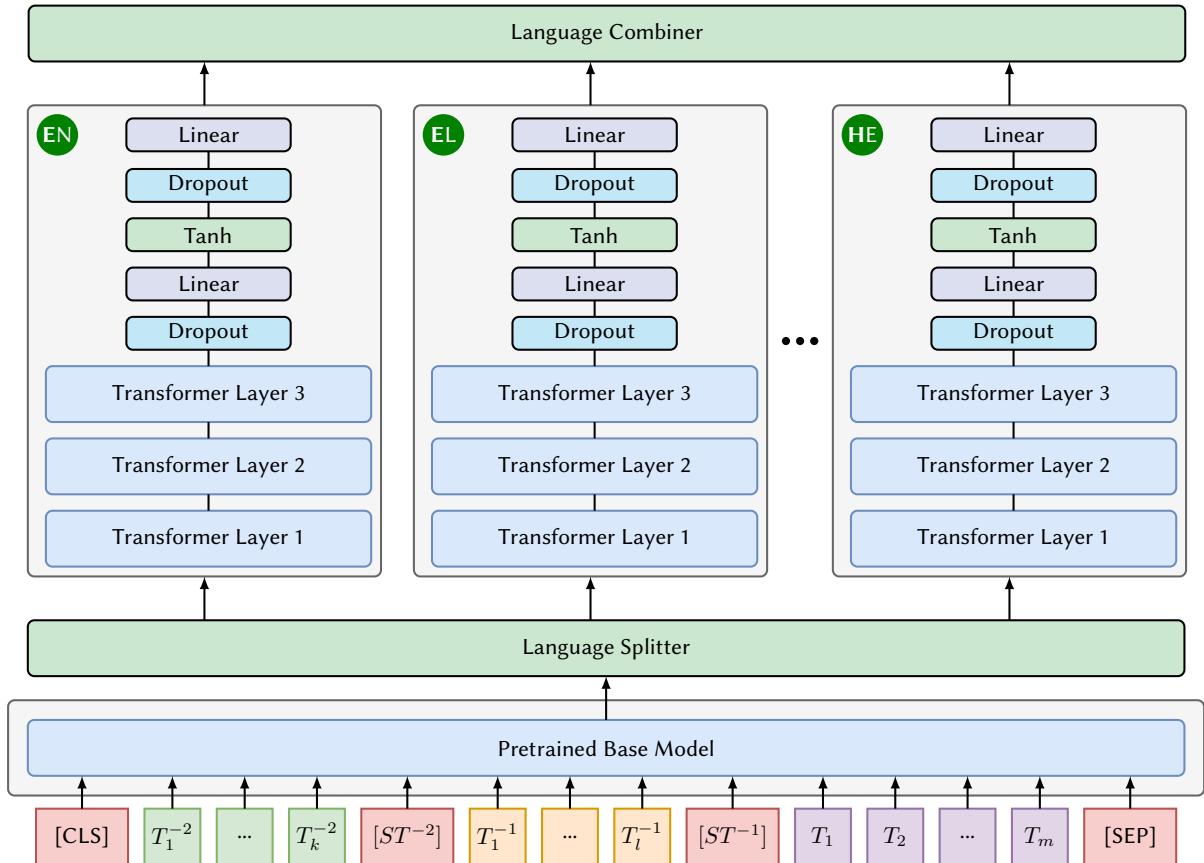


Figure 2: The figure presents the custom architecture that was used for the experiments. The model input consists of the [CLS] token, the tokens of the two preceding sentences (when available), (T_1^{-2} to T_k^{-2} and T_1^{-1} to T_k^{-1} , respectively) along with their corresponding special tokens ($[ST^{-2}]$ and $[ST^{-1}]$, respectively), and the tokens of the sentence to be classified (T_1 to T_m). The [ST] token is a special token that corresponds to one of the 19 (or 38) classes. Each custom Transformer head corresponding to each language had three Transformer Layers for RoBERTa-large and the corresponding XLM variant, one Transformer layer for XLM-RoBERTa-xlarge, while DeBERTa-v2-xxl did not have any Transformer layers.

Each custom Transformer head comprises the following components:

1. Three Transformer Layers which incorporate:

- a) Self-Attention Mechanism: Allows the model to focus on different parts of the input sequence.
 - b) Layer Normalization: Stabilizes and accelerates the training process.
 - c) Feed-Forward Neural Network: Introduces non-linearity and complexity.
 - d) Residual Connection: Helps in mitigating the vanishing gradient problem and allows deeper networks.
 - e) Dropout: Prevents overfitting by randomly dropping units during training.
2. Classification Process:
- a) The [CLS] token from the last Transformer layer (Transformer Layer 3) is passed through a dropout layer followed by a linear layer.
 - b) Finally, the output of the previous linear layer is passed through a Tanh activation function and then subjected to a dropout and a linear layer. The last linear layer produces logits corresponding to the number of classes.

Regarding the model training workflow, during each training iteration:

1. The input batch is fed into the pre-trained base model (Transformer).
2. The output of the pre-trained model is passed through the language splitter which splits it according to the language identifiers within the batch. Each split tensor is directed to the corresponding custom Transformer head based on its language for further processing.
3. The logits produced by each custom Transformer head are concatenated into a single batch through the language combiner.
4. The concatenated logits batch is passed through the loss function to compute the training loss.
5. Model performs backpropagation.

This approach allows the model to handle multiple languages effectively by utilizing specialized components tailored to the linguistic features and annotation styles of each language.

4.2. Experimental Setup

The decision to utilize open-source Transformer-based multi-lingual models, specifically obtained from the Hugging Face platform, in our research was motivated by the multi-lingual composition of the data, which encompassed nine distinct languages. These models have been pre-trained across various languages, rendering them an optimal choice for providing a robust and comprehensive framework for analyzing and interpreting multi-lingual data. Consequently, we employed such models to ensure the effective capture of language-specific nuances and contexts, leading to more accurate and reliable results. For the data that underwent automatic translation into English, we employed open-source Transformer-based models that were exclusively pre-trained in English. This approach ensured an optimal understanding and interpretation of the nuances of the English language, thereby bolstering the accuracy of our analysis.

We utilized the multi-lingual base version of the RoBERTa Transformer-based language model [22], XLM-RoBERTa-base⁴ [23], which underwent pre-training on 100 languages with 768 layers. This model was employed to conduct preliminary experiments for multi-label text classification using *AutoModelForSequenceClassification*. Subsequently, baseline scores were obtained during the exploratory phase (See section 3.3). After analyzing the baseline results for individual languages and all languages collectively (see Table 4 in Appendix A), we leveraged the larger 1024-layer version, XLM-RoBERTa-large⁵ [23], to conduct further experiments involving loss functions, class weights, and different class thresholds (see sections 4.3 and 4.4). The purpose was to address the challenges of class imbalance and language disparities. These experiments were primarily facilitated using the Transformers and Hugging Face libraries, in conjunction with 2 NVIDIA TITAN RTX GPU cards, with 24GB VRAM each.

⁴<https://huggingface.co/FacebookAI/xlm-roberta-base>

⁵<https://huggingface.co/FacebookAI/xlm-roberta-large>

For both sub-tasks, we fine-tuned two Transformer-based language models for the multi-lingual data using the custom model architecture with multiple heads presented in section 4.1, with each head focusing on a specific language. The employed models were the XLM-RoBERTa-large [23] and the XLM-RoBERTa-xl⁶ [24], with 1024 and 2560 layers, respectively. In the case of the English-translated data, we integrated the RoBERTa-large⁷ [22] and the DeBERTa-v2-xxl⁸ [25] models, consisting of 1024 and 1536 layers, respectively, into the custom multi-head architecture, focusing solely on English for both sub-tasks.

We initially fine-tuned our models using the provided training data and fine-tuned them using the validation set. During the fine-tuning process, we established the hyperparameters, finalized the loss function, and determined the best thresholds for our submitted results. Then, we combined the training and validation data to use as the training set for fine-tuning, without having a separate validation set, using the previously defined hyperparameters. An overview of the hyperparameters used for our experiments and submissions is provided in Table 5 of the Appendix A.

As for the custom model architecture, the custom head for RoBERTa-large and XLM-RoBERTa-large included three Transformer layers, while the custom head for XLM-RoBERTa-xl employed only one. Due to GPU VRAM memory and time limitations, the DeBERTa-v2-xxl did not incorporate any Transformer layers in its custom head. The experiments with the custom model architecture, which form the final submissions, were conducted using 2 NVIDIA H100 PCIe GPU cards, with 80GB VRAM each.

4.3. Loss Functions & Class Weights

Various loss functions, including Binary Cross-Entropy Loss with Logits⁹, Focal Loss [26], Class-balanced Loss [27], Distribution-Balanced Loss [28], and Class-balanced Negative Tolerant Regularization Loss [29], were tested by modifying the *Trainer* class from Hugging Face. These loss functions were employed as they were originally developed for handling data imbalance issues. They have previously been employed for the detection of human values by the PAI team in SemEval-2023 [15]. Positive weights were also calculated for each class to give more importance to the under-represented classes during model training, thereby improving the model’s performance in these classes. The experiments using the XLM-RoBERTa-large with the standard classification head (*AutoModelForSequenceClassification*) showed that the Binary Cross-Entropy Loss with Logits achieved the best results. Therefore, this loss function was used for all the submitted runs with and without class positive weights.

4.4. Thresholds

Initial experiments were conducted with various thresholds ranging from 0,1 to 0,95. The Macro-F1 score for all classes and each class separately was calculated during fine-tuning and evaluating with the provided validation set. After applying the sigmoid function to the validation and test set predictions, the predictions were converted into 1 if they were equal to or higher than the threshold and 0 if they were lower than the threshold. Consequently, 3 separate prediction files were created based on the 0,5 default threshold, the best general threshold for all classes, and the best threshold for each class. Based on the results from the validation set, the prediction file utilizing the optimal threshold for each class demonstrated the highest scores. Therefore, all predictions submitted for the test set were generated by determining the optimal threshold for each class individually.

⁶<https://huggingface.co/facebook/xlm-roberta-xl>

⁷<https://huggingface.co/FacebookAI/roberta-large>

⁸<https://huggingface.co/microsoft/deberta-v2-xxlarge>

⁹https://pytorch.org/docs/stable/generated/torch.nn.functional.binary_cross_entropy_with_logits.html

5. Results

5.1. Sub-task 1

The data presented in Table 2 illustrates that our test set submissions demonstrated significant improvement over the baseline scores in both the multi-lingual and translated English datasets. Among the multi-lingual multi-label multi-head models, the XLM-RoBERTa-xl model achieved the highest Macro-F1 score (39%) across all 38 classes by utilizing context and special tokens without class positive weights and being fine-tuned on the combined training and validation data as the training set. Conversely, the XLM-RoBERTa-large model, employing context, special tokens, class positive weights, and fine-tuned on 19 classes using only the training data, achieved the lowest score (34%).

In the context of the translated English data, the XLM-RoBERTa-large model, utilizing context and special tokens without class positive weights and having been fine-tuned on the combined training and validation data as the training set, produced the lowest Macro-F1 score across all classes (35%). At the same time, the remaining submissions yielded identical scores (37%).

Upon examining the F1 scores for each class individually, it becomes apparent that the *Universalism: nature* class achieved the highest F1 score at 63%, signifying successful detection by the models, as the remaining scores do not fall below 59%. Conversely, the classes with lower frequency in the texts were less accurately detected by the models. For instance, values such as *Humility* received a 0% F1 score in most submissions, with the highest score reaching only 11%. Furthermore, the models struggled to accurately classify the *Self-direction: thought* value, as their scores remained below 20%. Despite being one of the minority classes, the models correctly detected at least 27% of the annotated labels in the *Universalism: tolerance* class. The different model performance in classes is also evident in Figure 3 of the Appendix A, which illustrates the radar plot of the 19 values through the performance of our top-performing XLM-RoBERTa-xl model compared to the baseline models.

Table 2

Achieved F₁-score of each submission on the test dataset for sub-task 1. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

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: interpersonal	Rules	Humility	Beneficence: caring	Dependability	Universalism: concern	Universalism: nature	Universalism: tolerance
multi-lingual XLM-RoBERTa-large_weights_context_special tokens_19_only train data		34	13	20	28	28	37	37	45	22	33	46	46	49	21	04	32	32	47	63	21
multi-lingual XLM-RoBERTa-large_context_19		36	15	28	35	35	44	39	47	28	40	48	49	50	20	08	33	32	47	60	24
multi-lingual XLM-RoBERTa-xl_context_special tokens_19		38	15	27	31	36	43	41	51	32	44	49	48	51	23	00	34	35	50	63	24
multi-lingual XLM-RoBERTa-xl_context_special tokens_38		39	15	27	30	37	45	42	49	31	42	49	46	51	24	00	34	33	47	63	27
translated XLM-RoBERTa-large_context_special tokens_19	✓	35	14	25	30	28	41	40	46	25	40	48	48	48	20	05	34	30	46	59	25
translated RoBERTa-large_weights_context_special tokens_19_only train data	✓	37	19	23	31	32	40	41	45	31	43	48	51	48	26	11	34	33	48	60	27
translated RoBERTa-large_context_special tokens_19	✓	37	16	28	33	35	43	38	48	28	44	48	51	49	27	05	34	27	48	61	27
translated DeBERTa-v2-xxl_context_special tokens_19_only train data	✓	37	15	26	32	32	44	40	45	32	41	47	49	50	24	05	34	33	48	62	27
translated RoBERTa-large_context_special tokens_38	✓	37	12	24	32	36	42	39	46	28	43	47	49	49	22	00	34	32	47	61	27
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

5.2. Sub-task 2

The data presented in Table 3 illustrates that our submission for the multi-lingual test dataset outperformed the baseline score. Utilizing the XLM-RoBERTa-xl for the multi-lingual dataset and the

Table 3

Achieved F₁-score of each submission on the test dataset for sub-task 2. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

Submission	EN	All	F ₁ -score																		
			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			
multi-lingual XLM-RoBERTa-xl_context_special tokens_38		77	73	73	77	75	78	77	79	71	78	79	77	78	74	25	74	77	78	84	71
translated RoBERTa-large_context_special tokens_38	✓	77	72	72	78	74	78	78	78	73	78	78	78	77	73	22	78	77	78	82	74
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

RoBERTa-large for the English dataset, both leveraging context and special tokens without class positive weights and having been fine-tuned on the combined training and validation data as the training set, resulted in identical Macro-F1 scores across all 38 classes (77%). Once again, the class with the lowest F1 score was *Humility*, scoring 25% and 22% in the multi-lingual and English test datasets, respectively, significantly lower than the baselines' scores. Conversely, the *Universalism: nature* value yielded the highest F1 scores in both of our submissions. Finally, the *Universalism: tolerance* value was once again successfully detected by the models, despite being underrepresented in the data, achieving 71% and 74% in the multi-lingual and English test datasets, respectively.

6. Conclusion & Future Work

Our system, developed for Touché at CLEF 2024, addressed the "Human Value Detection" shared task by participating in both sub-tasks. This involved the fine-tuning of four Transformer language models within a custom multi-head model architecture for multi-label text classification. Our experimental approach encompassed the utilization of loss functions and class positive weights, as well as the incorporation of previous sentences as context and the representation of human values as special tokens. These measures were implemented to mitigate class imbalance and enhance the models' capacity to comprehend and classify texts more effectively.

Our submissions demonstrated superior performance compared to the baseline and other participating teams' scores in both the multi-lingual and English-translated test datasets, resulting in achieving the 1st place in sub-task 1. Despite scoring lower than the baseline in sub-task 2 in the English test dataset, our submission for the multilingual test dataset surpassed the baseline score. Notably, the XLM-RoBERTa-xl model, leveraging context and special tokens without class positive weights and fine-tuned on the combined training and validation data, exhibited strong performance in both sub-tasks for the multilingual data. Furthermore, our findings indicated that while class positive weights augmented the models' ability to classify under-represented classes, they did not yield an overall performance improvement. The shared task posed a significant challenge due to the presence of data imbalance across classes and languages, as well as the existence of low-resource languages in the texts.

To further optimize model performance for multi-label human value detection, future endeavors should center on exploring additional Transformer layers within the custom multi-head architecture, with a particular emphasis on even larger Transformer language models such as the XLM-RoBERTa-

xxl¹⁰. Additionally, the investigation of alternative loss functions to address data imbalance, the implementation of data augmentation methods or even an ensemble of various models hold the potential to further enhance performance.

7. Limitations

The experimentation process in both sub-tasks has revealed a significant issue of class imbalance. Despite the assignment of higher weights to the minority classes, it has become evident that detecting one or more human values is a challenging task. This challenge primarily stems from the imbalance in the annotated human values across languages as well as the general class imbalance among human values in the multi-lingual training dataset. Moreover, the presence of low-resource languages such as Hebrew and Greek has posed a further challenge, as the multi-lingual models contain a smaller number of tokens for these languages in comparison to English. Notwithstanding these challenges, the multi-lingual models have performed adequately compared to the baseline models. Moreover, in the process of fine-tuning the XLM-RoBERTa-xl and DeBERTa-v2-xxl models, we encountered challenges stemming from limitations in GPU VRAM memory and time. Specifically, we modified the fine-tuning approach for the first model by reducing the number of Transformer layers from three to one. Furthermore, in the second model's case, the custom head's multi-head architecture did not incorporate any Transformer layers.

Acknowledgments

The research leading to these results has received funding from the European Union’s Horizon Europe research and innovation programme, in the context of: TITAN project, under grant agreement No. 101070658 and AI4TRUST project, under grant agreement No. 101070190. This paper reflects only the view of the authors and the European Commission is not responsible for any use that may be made of the information it contains.

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). doi:10.1037/a0029393.
- [2] D. Narvaez, J. Rest, The four components of acting morally, *Moral behavior and moral development: An introduction* 1 (1995) 385–400.
- [3] M. Ruskov, M. Dagioglou, M. Kokol, S. Montanelli, G. Petasis, et al., A knowledge graph of values across space and time, in: *CEUR Workshop Proceedings*, volume 3536, CEUR-WS, 2023, pp. 8–20.
- [4] M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, L. Scheunemann, *Values and Identities - a Policymaker’s Guide*, Technical Report KJ-NA-30800-EN-N, European Commission’s Joint Research Centre, Luxembourg, 2021. doi:10.2760/349527.
- [5] E. Aharoni, S. Fernandes, D. J. Brady, C. Alexander, M. Criner, K. Queen, J. Rando, E. Nahmias, V. Crespo, Attributions toward artificial agents in a modified moral turing test, *Scientific Reports* 14 (2024) 8458.
- [6] 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*

¹⁰<https://huggingface.co/facebook/xlm-roberta-xxl>

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.

- [7] G. Papadopoulos, M. Kokol, M. Dagioglou, G. Petasis, Andronicus of rhodes at SemEval-2023 task 4: Transformer-based human value detection using four different neural network architectures, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 542–548. URL: <https://aclanthology.org/2023.semeval-1.75>. doi:10.18653/v1/2023.semeval-1.75.
- [8] A. F. Ntogramatzis, A. Gradou, G. Petasis, M. Kokol, The ellogen web annotation tool: Annotating moral values and arguments, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3442–3450.
- [9] 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.
- [10] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, B. Stein, Identifying the human values behind arguments, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4459–4471. URL: <https://aclanthology.org/2022.acl-long.306>. doi:10.18653/v1/2022.acl-long.306.
- [11] S. Yu, D. Liu, W. Zhu, Y. Zhang, S. Zhao, Attention-based lstm, gru and cnn for short text classification, *J. Intell. Fuzzy Syst.* 39 (2020) 333–340. URL: <https://doi.org/10.3233/JIFS-191171>. doi:10.3233/JIFS-191171.
- [12] D. Cortiz, Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra, 2021. arXiv:2104.02041.
- [13] D. S. Brown, J. Schneider, A. D. Dragan, S. Niekuum, Value alignment verification, 2021. arXiv:2012.01557.
- [14] G. Balikas, John-arthur at SemEval-2023 task 4: Fine-tuning large language models for arguments classification, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1428–1432. URL: <https://aclanthology.org/2023.semeval-1.197>. doi:10.18653/v1/2023.semeval-1.197.
- [15] L. Ma, Z. Sun, J. Jiang, X. Li, PAI at SemEval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 256–261. URL: <https://aclanthology.org/2023.semeval-1.34>. doi:10.18653/v1/2023.semeval-1.34.
- [16] D. Schroter, D. Dementieva, G. Groh, Adam-smith at SemEval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 532–541. URL: <https://aclanthology.org/2023.semeval-1.74>. doi:10.18653/v1/2023.semeval-1.74.
- [17] C. Zhang, P. Liu, Z. Xiao, H. Fei, Mao-zedong at SemEval-2023 task 4: Label representation multi-head attention model with contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 426–432. URL: <https://aclanthology.org/2023.semeval-1.58>. doi:10.18653/v1/2023.semeval-1.58.

- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. [arXiv:1706.03762](#).
- [19] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goerriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:[10.1007/978-3-031-28241-6_20](#).
- [20] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, B. Stein, Identifying the Human Values behind Arguments, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Association for Computational Linguistics, 2022, pp. 4459–4471. doi:[10.18653/v1/2022.acl-long.306](#).
- [21] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, B. Valentin, D. Dastgheib, O. Ghahroodi, M. A. Sadraei, E. Asgari, L. Kawaletz, H. Wachsmuth, B. Stein, The touché23-valueeval dataset for identifying human values behind arguments, 2023. URL: <https://arxiv.org/abs/2301.13771>. [arXiv:2301.13771](#).
- [22] 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](#).
- [23] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. [arXiv:1911.02116](#).
- [24] N. Goyal, J. Du, M. Ott, G. Anantharaman, A. Conneau, Larger-scale transformers for multilingual masked language modeling, 2021. [arXiv:2105.00572](#).
- [25] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. [arXiv:2006.03654](#).
- [26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: <https://arxiv.org/abs/1708.02002>. [arXiv:1708.02002](#).
- [27] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, 2019. URL: <https://arxiv.org/abs/1901.05555>. [arXiv:1901.05555](#).
- [28] T. Wu, Q. Huang, Z. Liu, Y. Wang, D. Lin, Distribution-balanced loss for multi-label classification in long-tailed datasets, 2021. URL: <https://arxiv.org/abs/2007.09654>. [arXiv:2007.09654](#).
- [29] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, E. Ozkirimli, Balancing methods for multi-label text classification with long-tailed class distribution, 2021. URL: <https://arxiv.org/abs/2109.04712>. [arXiv:2109.04712](#).

A. Appendix

Table 4

Achieved F₁-score of the baseline XLM-RoBERTa-base on the validation dataset for sub-task 1. The model was fine-tuned using the whole multi-lingual training dataset and evaluated in multi-lingual validation dataset.

	All	Languages								
		English	Greek	German	French	Italian	Dutch	Bulgarian	Turkish	Hebrew
XLM-RoBERTa-base	29,5	22,41	26,16	25,24	2,52	22,71	18,71	23,30	28,03	24,16

Table 5

Models' Hyperparameters used for experiments and submissions. For the XLM-RoBERTa-xl model, 4 batch size was used for training, validation and testing as well as bf16 for mixed precision training. For the DeBERTa-v2-xxl model, 4 batch size was used for training, validation and testing as well as fp16 for mixed precision training.

Hyperparameter	Value
Seed	2024
Number of Epochs	20
Early Stopping Patience	5
Sequence Length	512
Train Batch Size	8 / 4
Validation / Test Batch Size	8 / 4
Learning Rate	5e-6
Weight Decay	0.01
Warm-up Ratio	0.01
Optimizer	AdamW
AdamW Epsilon	1e-8
LR Scheduler	Linear
Mixed Precision	fp16 / bf16

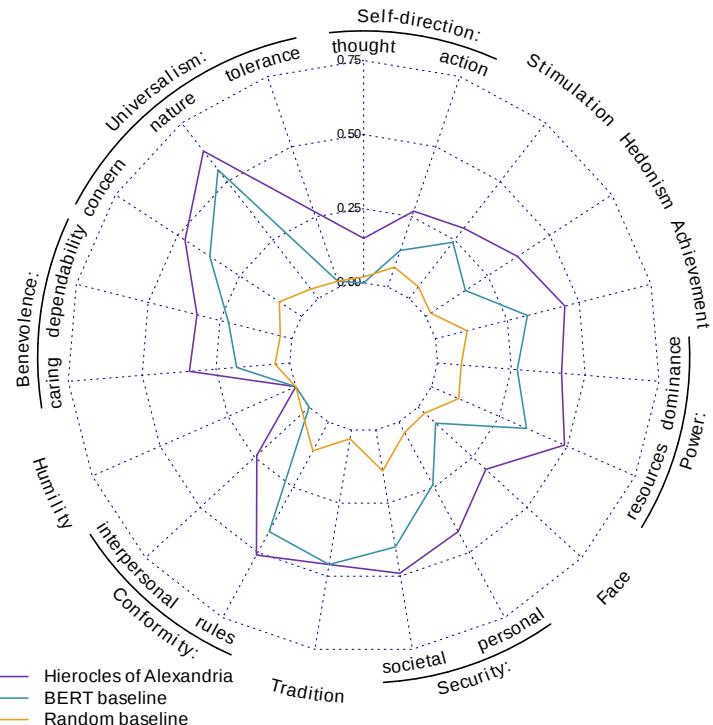


Figure 3: Radar Plot of 19 value categories (sub-task 1) illustrating the different performance per value in F1 score of our best-performing model compared to the baselines.