Touché at CLEF 2024 | Human Value Detection

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

Sotirios Legkas, Christina Christodoulou, Matthaios Zidianakis, Dimitrios Koutrintzes, Maria Dagioglou, and Georgios Petasis

DEMOKRITOS

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

OVERVIEW

HUMAN VALUE DETECTION@SEMEVAL23

- Dataset: Arguments
  - Premise, Conclusion, Stance
  - Monolingual task (English)

- Our approach: Multi-task ensemble Model architecture
  - Main motive: handle class imbalance

HUMAN VALUE DETECTION@CLEF24

- Dataset: Texts (400-800 words)
  - Multilingual task (9 languages + English translations)

- Our approach: Multi-task Model architecture
  - *Challenge 1:* Handle class imbalance
  - *Challenge 2:* Handle multiple languages
  - *Challenge 3:* Exploit context

EXPLORATORY PHASE (1 / 2)

Empirical Evidence (XLM-RoBERTa, Conneau et al., 2020[1]):

- Superior performance when fine-tuned with multilingual data
    - Our work is compliant to the empirical results

Initial experiment:

- Fine-tuned XLM-RoBERTa model:
    - Single model trained on all available languages
    - Multiple models each for a single language

[1] 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.

EXPLORATORY PHASE (2 / 2)

Observations:

- Models fine-tuned with multilingual data outperform monolingual ones
  - Validates empirical evidence
- Performance ***varies significantly*** across languages

Variation across languages can be attributed to:

- Language disparities
- Class imbalance across languages

	macro-F1 (XLM-RoBERTa) (base)
All	29.5
English	22.41
Greek	26.16
German	25.24
French	2.52
Italian	22.71
Dutch	18.71
Bulgarian	23.30
Turkish	28.03
Hebrew	24.16

EXPLORATORY PHASE (2 / 2)

<u>Observations:</u>

- Models fine-tuned with multilingual data outperform monolingual ones
    - Validates empirical evidence
- Performance ***varies significantly*** across languages

<u>Variation across languages can be attributed to:</u>

- Language disparities
- Class imbalance across languages

	macro-F1 (XLM-RoBERTa) (base)
All	29.5
English	22.41
Greek	26.16
German	25.24
French	2.52
Italian	22.71
Dutch	18.71
Bulgarian	23.30
Turkish	28.03
Hebrew	24.16

EXPLORATORY PHASE (2 / 2)

Observations:

- Models fine-tuned with multilingual data outperform monolingual ones
  - Validates empirical evidence
- Performance ***varies significantly*** across languages

Variation across languages can be attributed to:

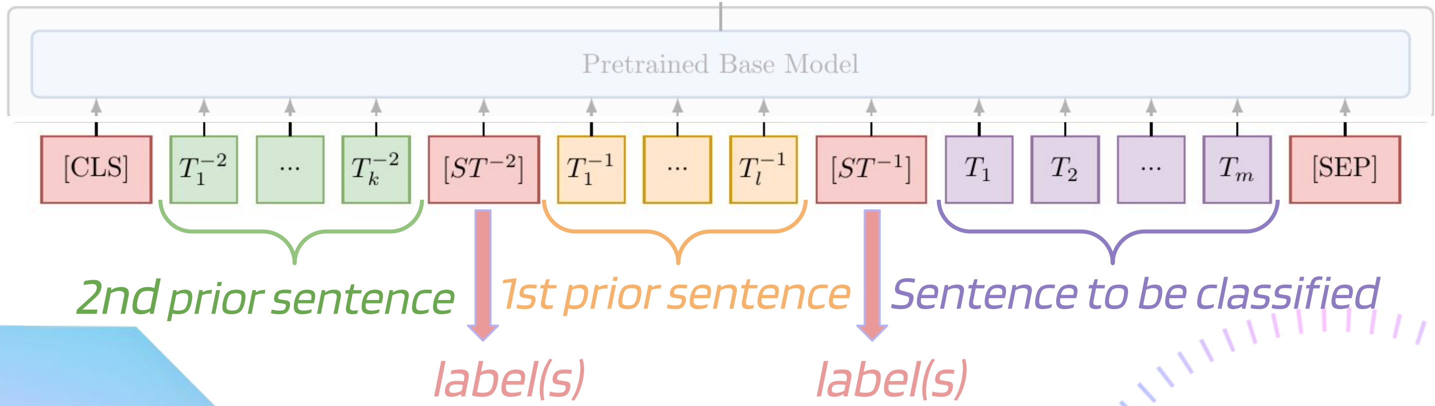
- Language disparities
- Class imbalance across languages

	macro-F1 (XLM-RoBERTa) (base)
All	29.5
English	22.41
Greek	26.16
German	25.24
French	2.52
Italian	22.71
Dutch	18.71
Bulgarian	23.30
Turkish	28.03
Hebrew	24.16

PROPOSED APPROACH: MODEL INPUT

Takes advantage of the available ***contextual information:***

- Sentence under examination is prepended with the history of the 2 previous sentences
    - Depending on sentence availability and model input capacity
- Added special tokens to the preceding sentences:
    - *Training:* The annotated values of these sentences (19/38 classes)
    - *Inference:* The previously predicted values of these sentences (19/38 classes)



PROPOSED APPROACH: MODEL ARCHITECTURE (1 / 7)

Considering:

- Multi-label classification task
- The language disparities
  - The linguistic nuances

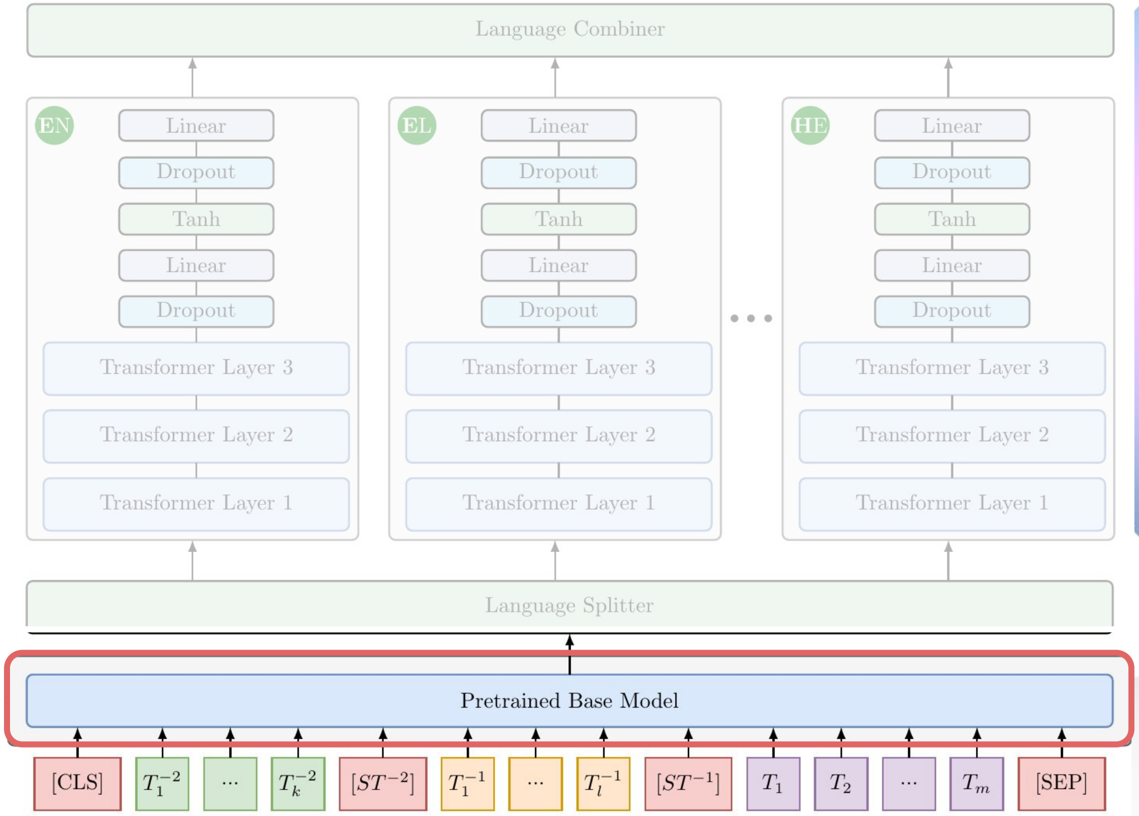
Our proposed approach:

- Multi-task learning
  - Each language is being considered as a separate task
- Multi-head architecture
  - Each task corresponds to a single head
- Model extended with custom classification heads

PROPOSED APPROACH: MODEL ARCHITECTURE (2 / 7)

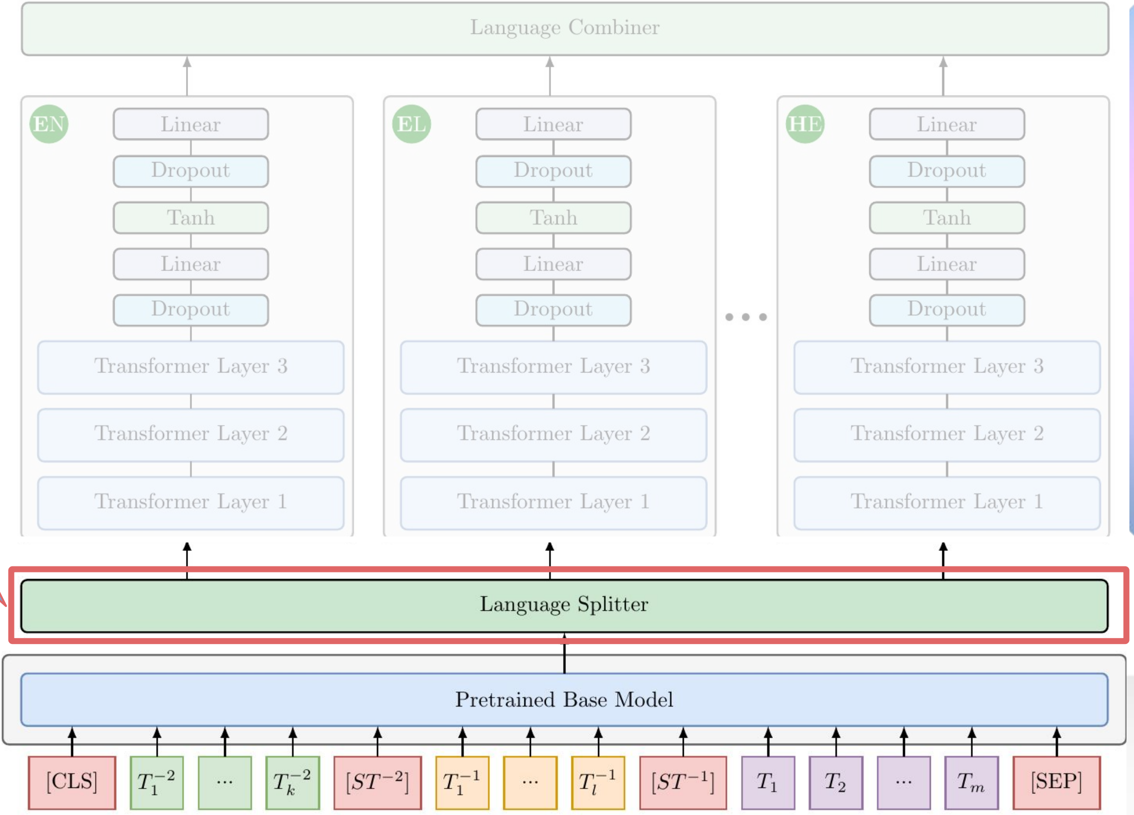
Foundation: Pre-trained Transformer language model (encoder)

- The input batch is fed into the pre-trained base model



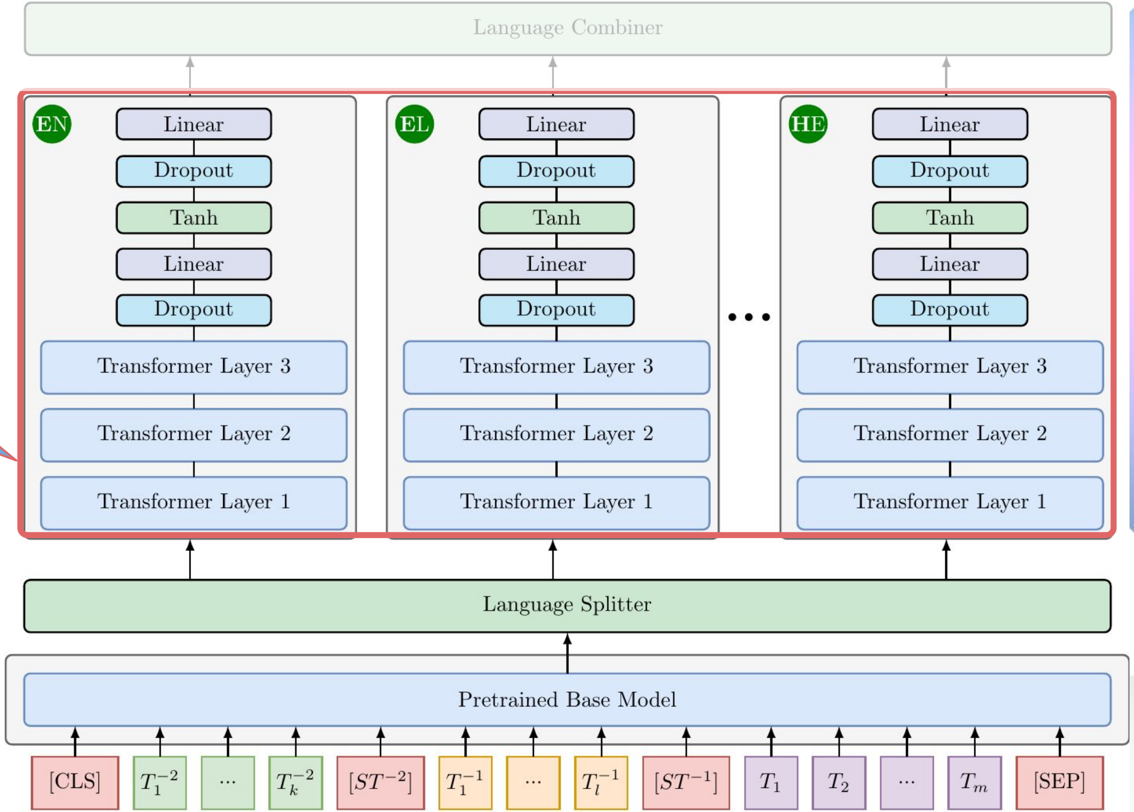
PROPOSED APPROACH: MODEL ARCHITECTURE (3 / 7)

Language Splitter: directs each tensor to its corresponding language-specific head



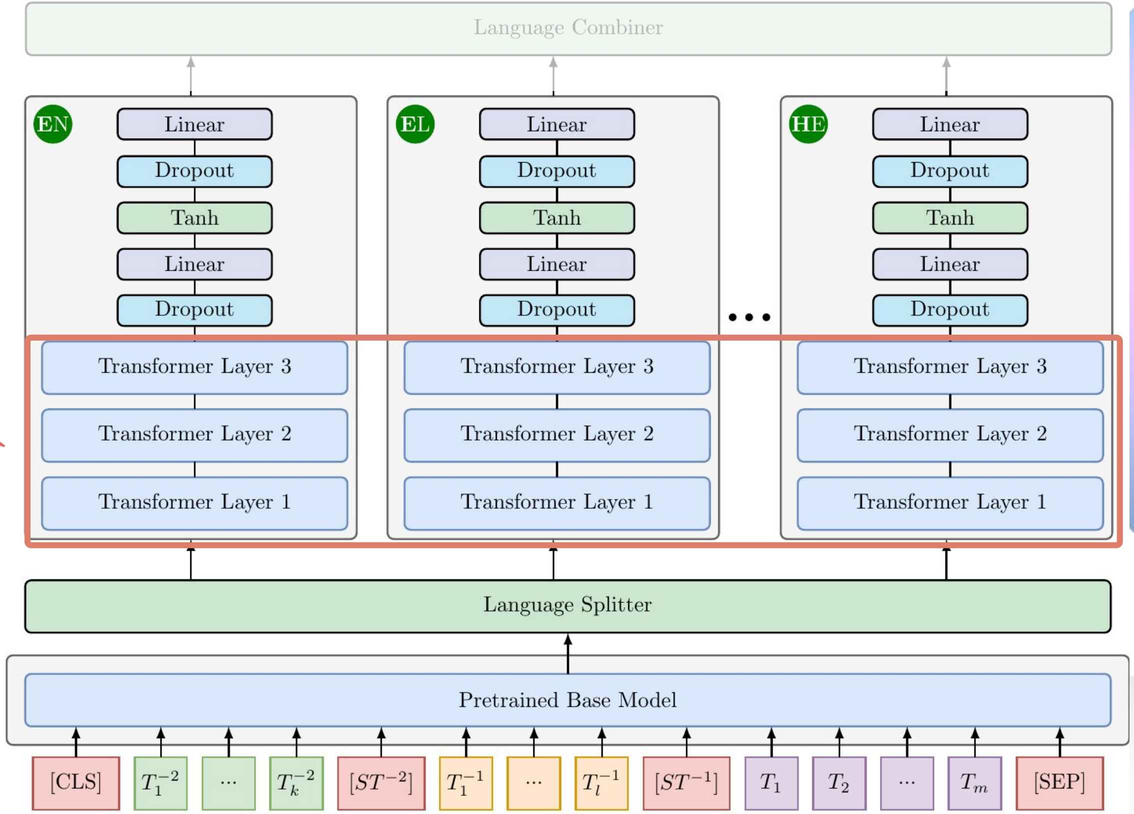
PROPOSED APPROACH: MODEL ARCHITECTURE (4 / 7)

9 custom heads added on top, each one for a specific language



PROPOSED APPROACH: MODEL ARCHITECTURE (5 / 7)

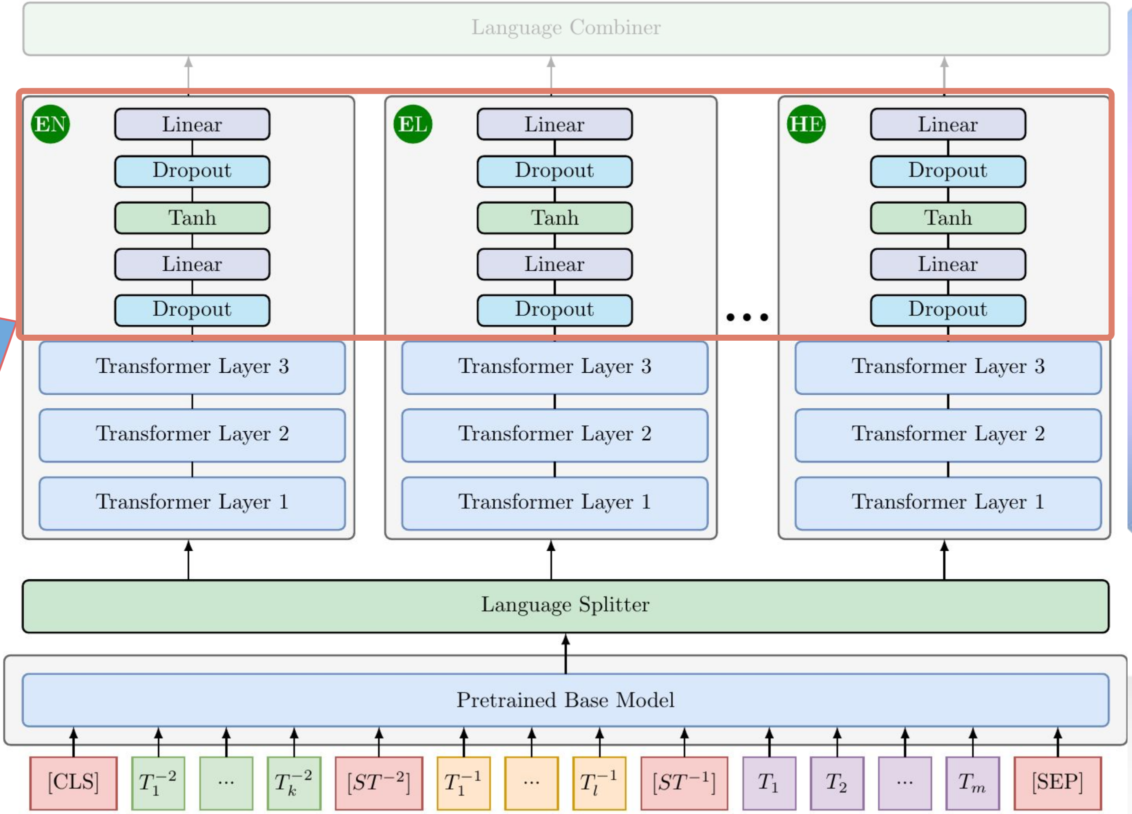
Each language-specific head comprises:
3 Transformer layers



PROPOSED APPROACH: MODEL ARCHITECTURE (6 / 7)

3rd transformer layer's **[CLS]** followed by:

- Dropout
- Linear layer
- Tanh
- Dropout
- Linear layer



Problem

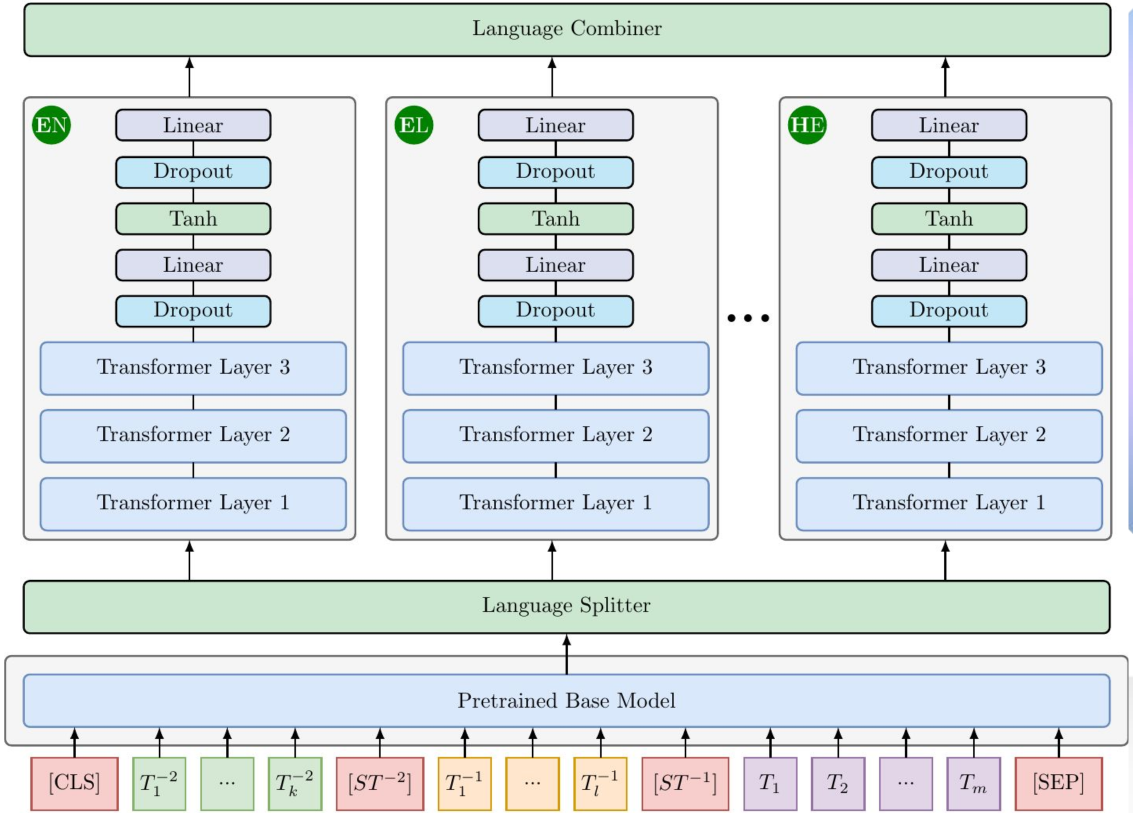
- Class imbalance → *unequal probabilities*

Solution: **?**

PROPOSED APPROACH: MODEL ARCHITECTURE (7 / 7)

Classification Thresholds:

- Thresholds per class
  - Extending last year's winning approach (Schroter et al., 2023[2])
- After sigmoid function applied to logits, predictions converted into:
  - **1**: if prediction >= threshold
  - **0**: if prediction < threshold



[2]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.

FINE-TUNING

Fine-tuning:

- Binary Cross-Entropy Loss with Logits achieved the best results
- Positive weights for each class (most for under-represented classes)
  - Only helpful within monolingual classifiers

Threshold calculation:

- Keep threshold that maximizes the macro-F1 per class
  - Threshold range [0.05, 0.95]
- Generated predictions using the optimal threshold for each class

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

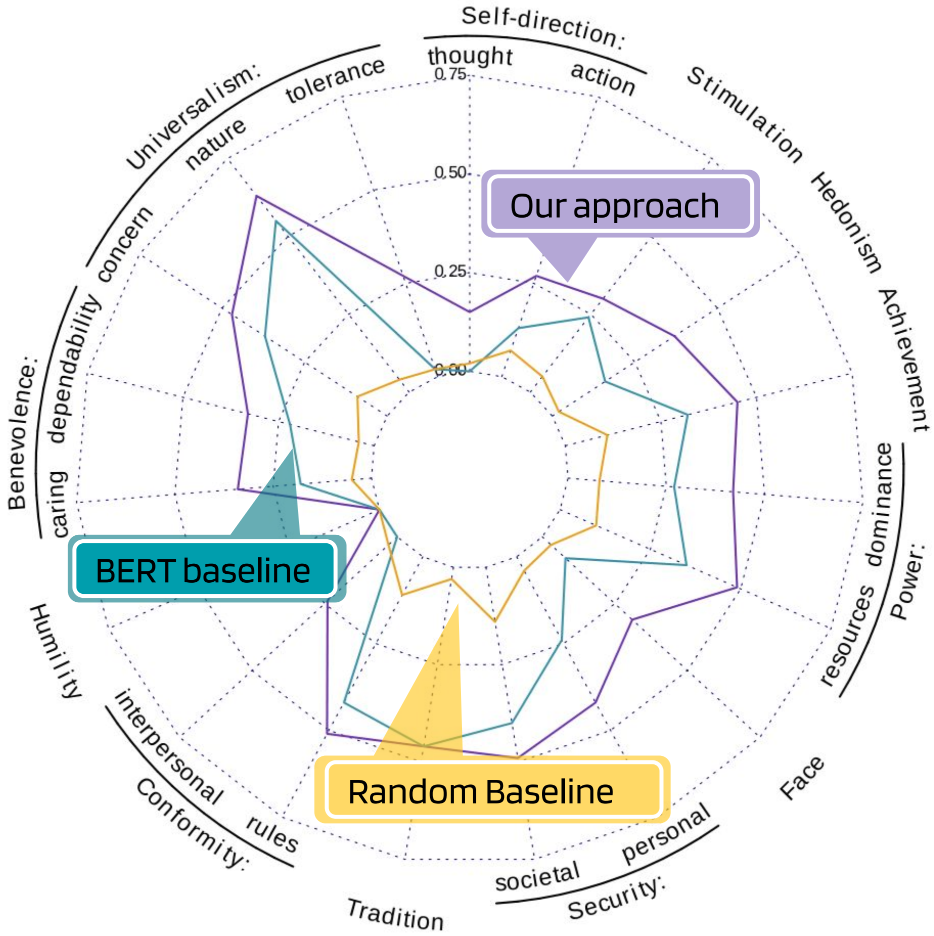
*Hardware: 1 NVIDIA H100 PCIe GPU cards, 80GB VRAM

RESULTS: SUB-TASK 1

- Our approach vs baselines (macro-F1):
  - Multilingual:
    - Custom XLM-RoBERTa-xl (**0.39**)
  - English:
    - Custom RoBERTa-large (0.37)
    - Custom DeBERTa-v2-xxl (0.37)

- Test set submissions outperformed baseline scores in both multilingual and English-translated datasets

- Our approach outperformed all other approaches for sub-task 1 in both multilingual and English-translated datasets



RESULTS: SUB-TASK 2

- Our approach vs baselines (macro-F1):
  - Multilingual: XLM-RoBERTa-xl
  - English: RoBERTa-large

- Outperforms all baselines
  - Except BERT-baseline (available only for English)

- Trained models with 38 classes to tackle both sub-task 1 & 2
  - Alternative solution: tackle sub-task 2 as a separate classification problem
    - Not tested due to competition time constraints

SUBMISSIONS	macro-F1 (multilingual)	macro-F1 (English)
Custom XLM-R-XL	0.77	-
Custom R-large	-	0.77
BERT-baseline	-	0.81
Random-baseline	0.53	0.53
Random-baseline (EN)	-	0.52

CONCLUSIONS

Key points:

- Multi-task Model architecture
  - Considered languages as separate tasks → Capture linguistic nuances and disparities
- Dealt with data imbalance using classification thresholds for each class
- Exploiting contextual information (previous sentences and their classification)

Achievements:

- 1st Place in sub-task 1 (Multilingual & English-translated datasets).
- Multilingual submission outperformed baseline in sub-task 2.

FUTURE WORK

- Experiment with Larger Models:
  - Add more Transformer layers within the custom architecture
  - Leverage as foundation larger models like XLM-RoBERTa-xxl
- Experiment with Data augmentation
- Experiment with Ensemble modeling
- Experiment with alternative loss functions
- Experiment different classification strategies
  - To better address sub-task 2

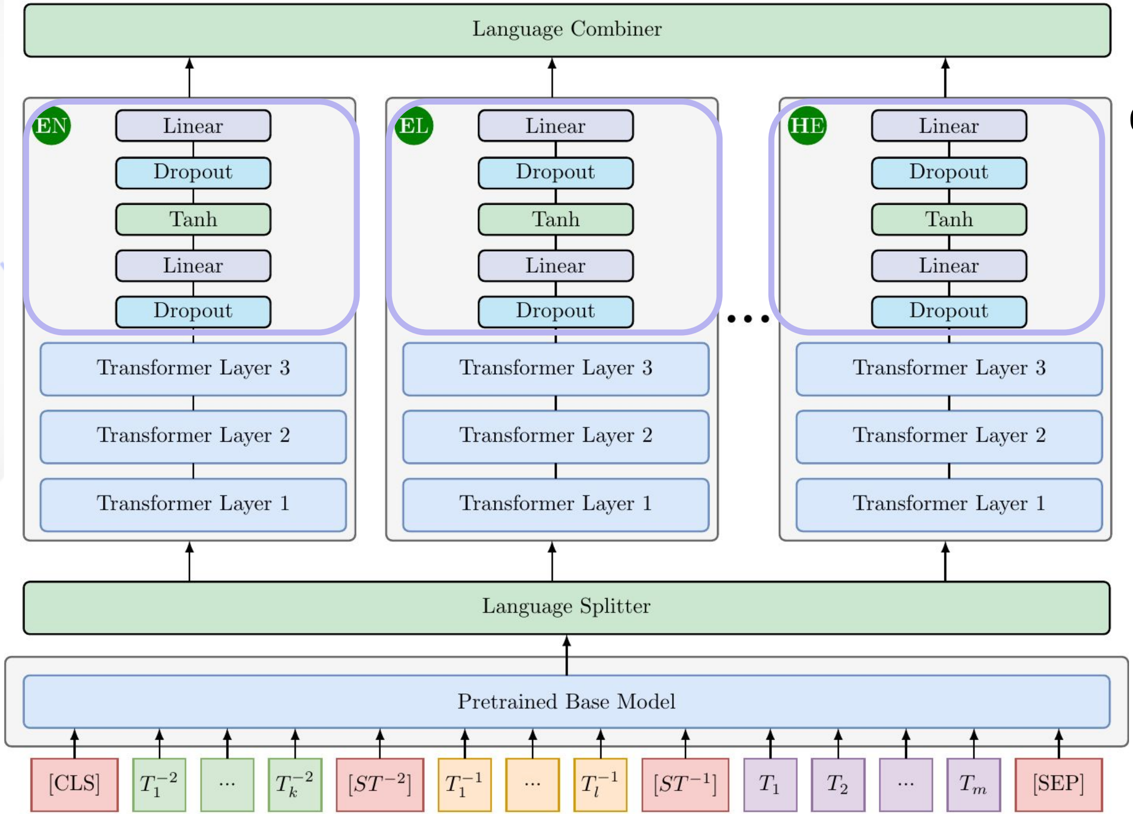
THANK YOU FOR YOUR ATTENTION!

Do you have any questions?

Sotiris Legkas

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

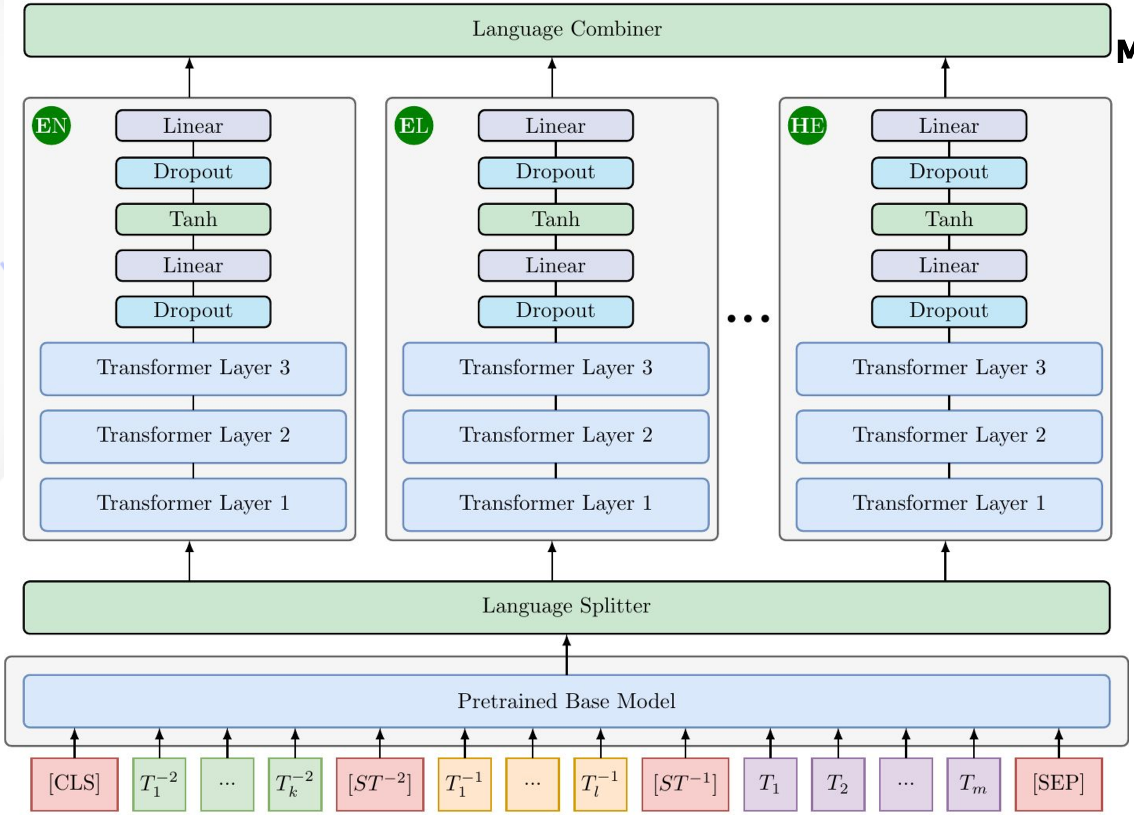
SYSTEM OVERVIEW: MODEL ARCHITECTURE



Classification Process:

1. The [**CLS**] token from the last Transformer layer (Transformer Layer 3) is passed through a **dropout** layer followed by a **linear** layer.
2. The output of the previous linear layer is passed through a **Tanh** activation function and then subjected to a dropout and a linear layer.
3. The last linear layer produces **logits** corresponding to the number of classes.

SYSTEM OVERVIEW: MODEL ARCHITECTURE



Model Training Workflow:

1. The **input batch** is fed into the pre-trained base model.
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**.

RESULTS: SUB-TASK 1

Achieved F_1-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: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: 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

RESULTS: SUB-TASK 2

Achieved F_1-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	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance
		F_1-score																			
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