Arthur Schopenhauer at Touché 2024: Multi-Lingual Text Classification Using Ensembles of Large Language Models

Hamza Yunis

CLEF 2024

Task Goal

- Identify human values that a text references (Subtask 1) and whether these values are attained or constrained (Subtask 2).
- Examples:

Text	**Referenced Values**
Widely considered one of the darkest days of the Troubles, relatives of the victims have met regularly to mourn their loss and campaign for justice.	Universalism: concern (attained)
We were hoping that we would get recourse to justice for our dead family members and that hasn't happened.	Universalism: concern (constrained)

Task Formulation

- Multi-label classification problem.
- Labeled dataset: 59662 texts from 9 languages.
- 38 label columns corresponding to 19 values from the Schwartz taxonomy (each attained or constrained).

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

Simplifying the Task

Preliminary Data Analysis

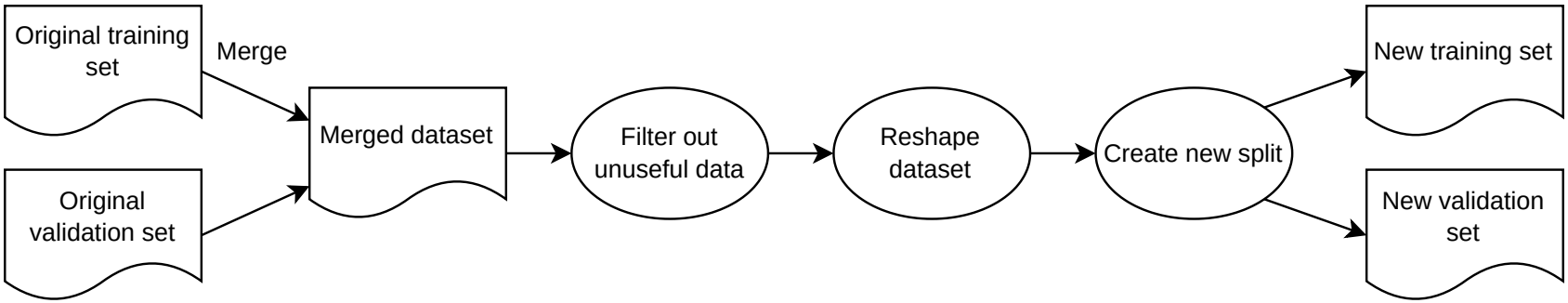
- 94% of texts have a single label or no label.
- Attainment / Constraint is largely independent from referenced values:

> We were hoping that we would get recourse to justice for our dead family members and that hasn't happened.

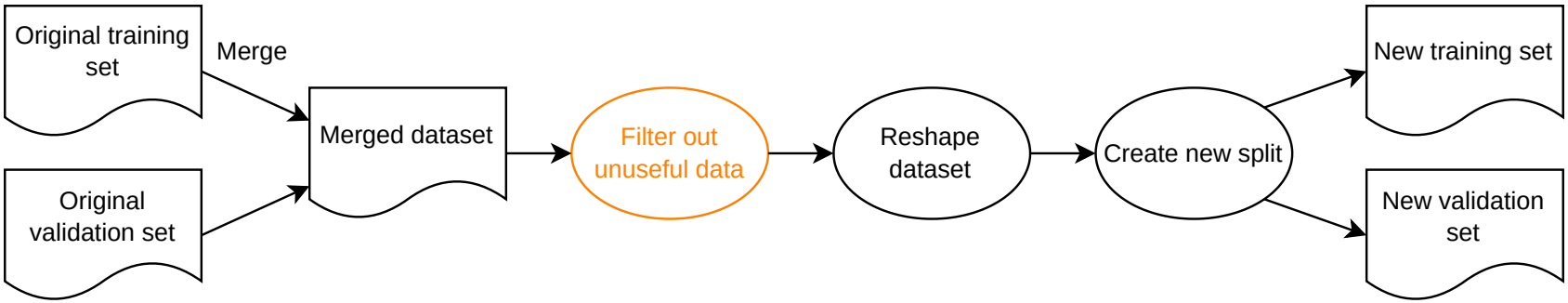
Resultant Simplification

- Restricting the training / predicting process in Subtask 1 to one label (including 'no label').
- Training the models that predict attainment independently from models that predict human values.

Data Preprocessing



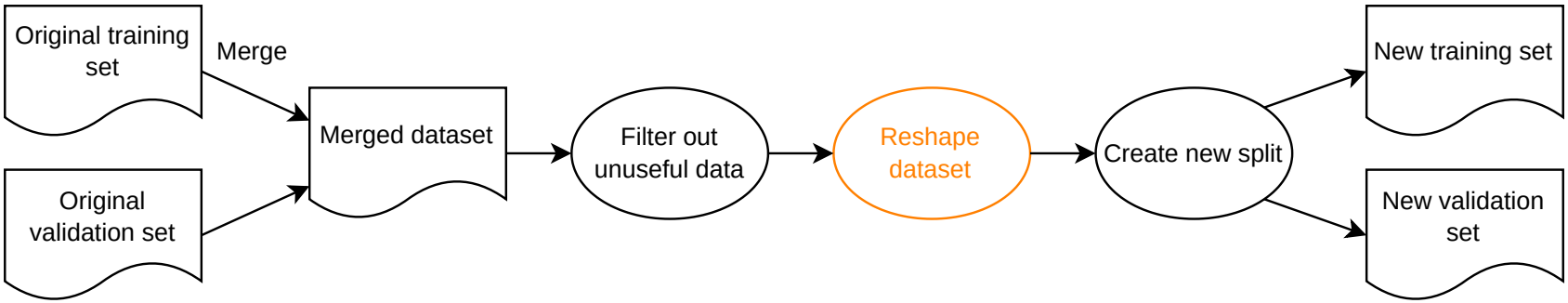
Data Preprocessing



- Duplicate texts.
- Texts with multiple labels.
- Texts with two words or less. Examples:

```
"76 Comments" "Extreme?" "It's Dr." "Moving out."
"PM" "Source:  PA." "Why?" "he said." "rise"
```

Data Preprocessing



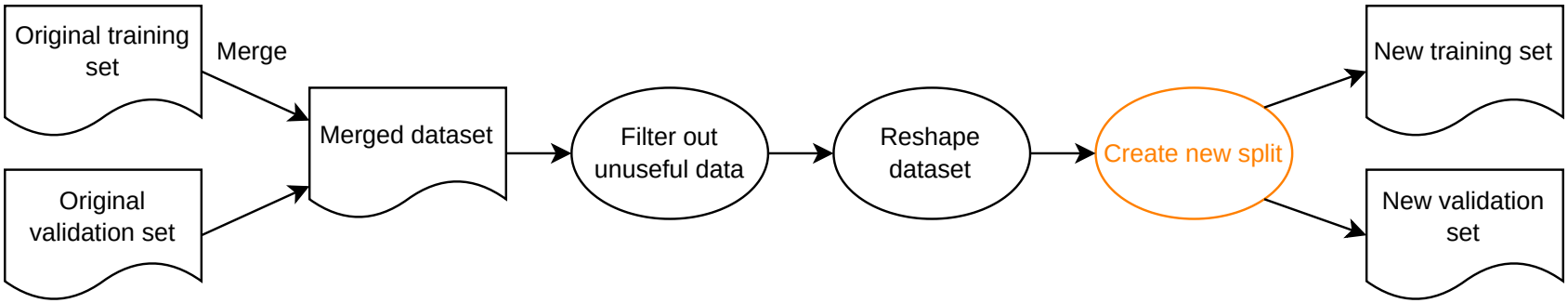
The 38 label columns are replaced by two columns:

hv_value Numeric code for the human value referenced by the text (including 'no label').

attainment Numeric code for attainment:

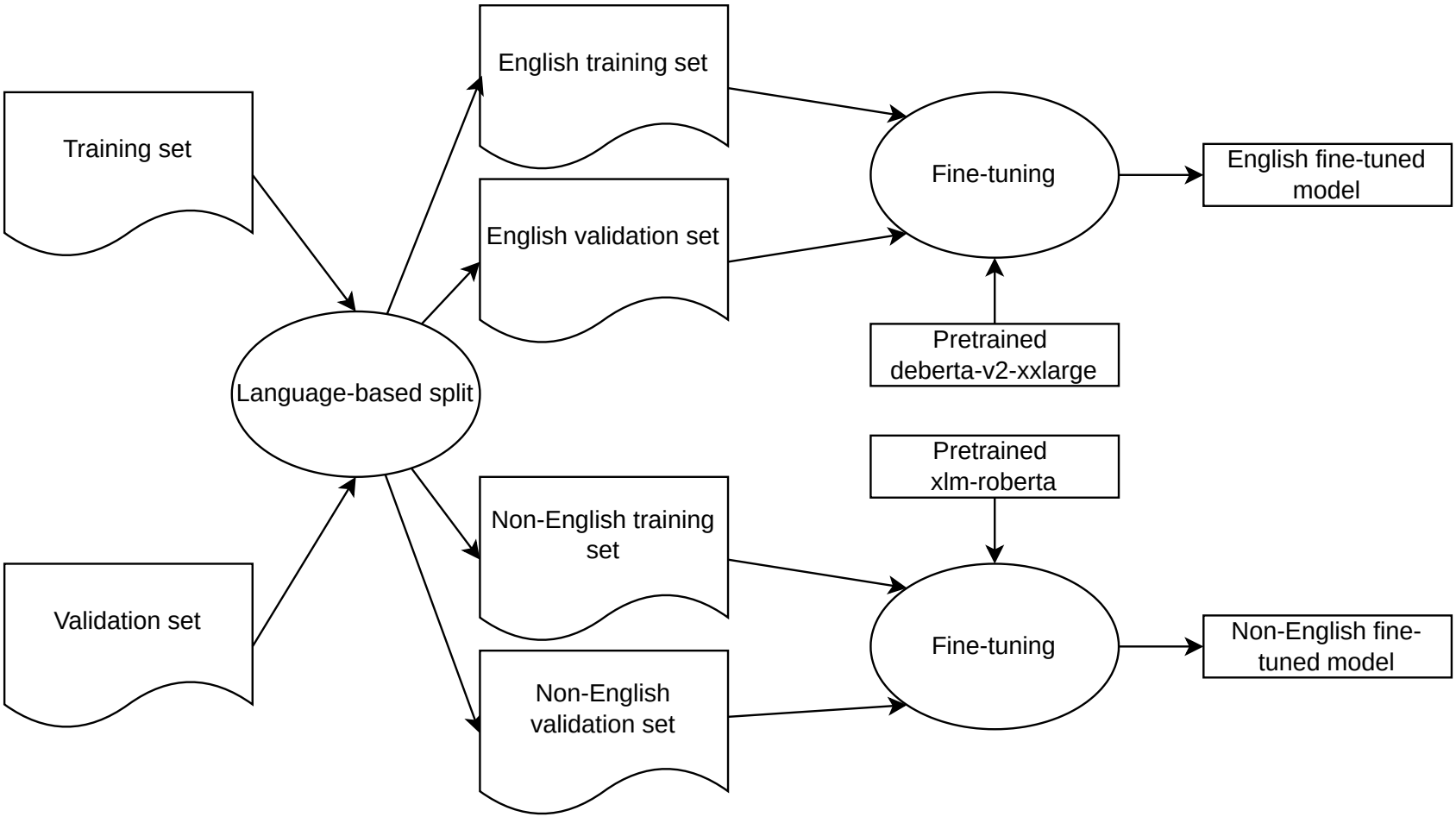
- 0: constrained
- 1: attained
- 2: NA

Data Preprocessing



- New validation set is created using 10% of the data.
- Proportional allocation is applied using language-label combinations.

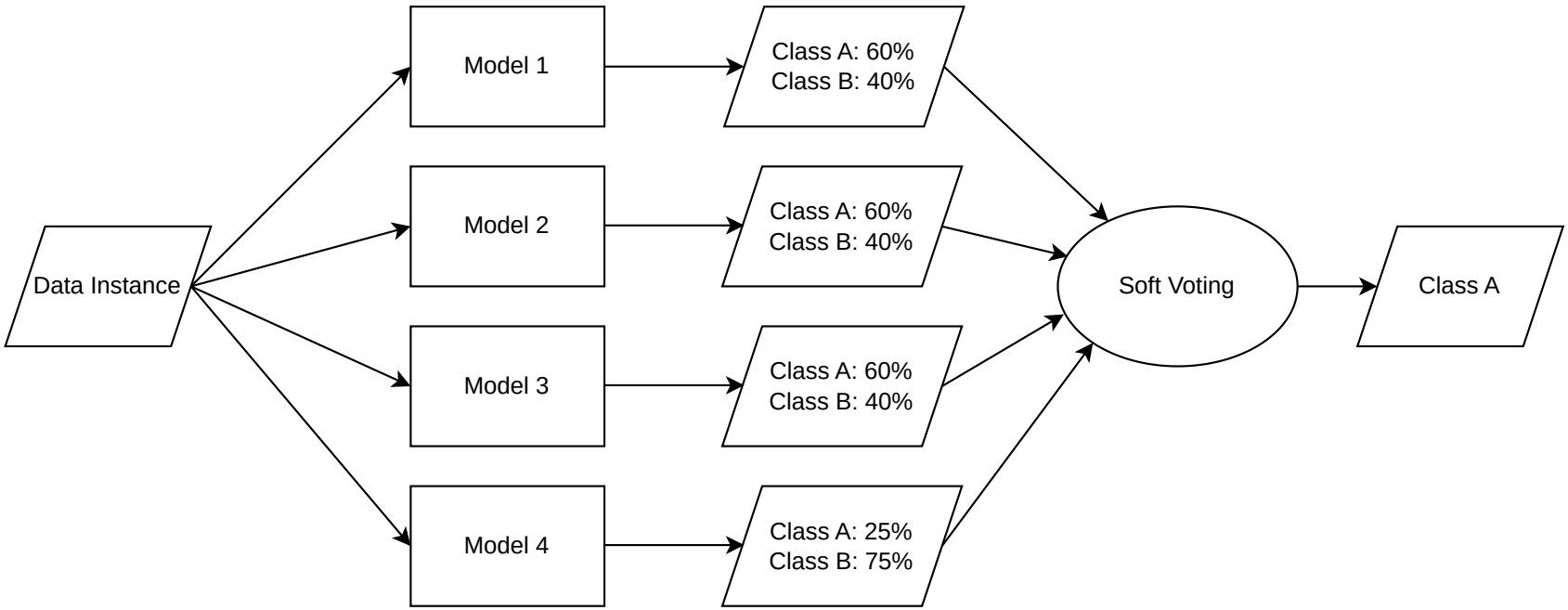
Fine-Tuning


Training set
Validation set
Language-based split
English training set
English validation set
Non-English training set
Non-English validation set
Fine-tuning
Fine-tuning
Pretrained deberta-v2-xxlarge
Pretrained xlm-roberta
English fine-tuned model
Non-English fine-tuned model

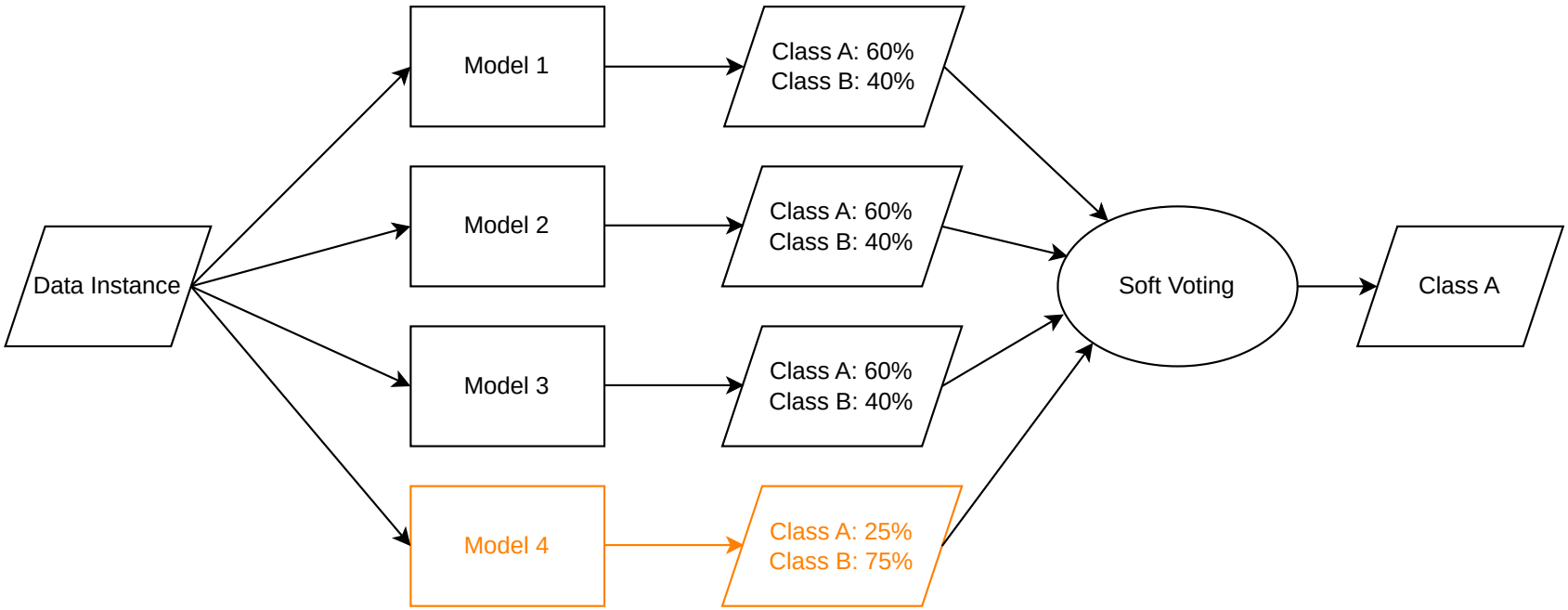
Model Overview

Model Name	Languages	Architecture	Seed	Loss Function
Subtask 1				
Model 1	English	deberta-v2-xxlarge	66	Cross-Entropy
Model 2	English	deberta-v2-xxlarge	66	Weighted Cross-Entropy
Model 3	English	deberta-v2-xxlarge	67	Cross-Entropy
Model 4	English	deberta-v2-xxlarge	67	Weighted Cross-Entropy
Model 5	Non-English	xlm-roberta	66	Cross-Entropy
Model 6	Non-English	xlm-roberta	66	Weighted Cross-Entropy
Model 7	Non-English	xlm-roberta	67	Cross-Entropy
Model 8	Non-English	xlm-roberta	67	Weighted Cross-Entropy
Subtask 2				
Model 9	English	deberta-v2-xxlarge	66	Cross-Entropy
Model 10	Non-English	xlm-roberta	66	Cross-Entropy

Pruned Soft Voting (Motivation)



Pruned Soft Voting (Motivation)



- Observation: Model 4's prediction has a significantly higher probability than the rest, indicating it is better trained for this data instance than the other models.
- Therefore: It is reasonable to adopt Model 4's prediction as the final prediction and neglect the remaining predictions.

Pruned Soft Voting

- General procedure: Given a threshold T, if there are predictions with probabilities exceeding T, then apply soft voting only to those predictions; otherwise, apply soft voting to all predictions.
- Finding the optimal threshold for Subtask 1 was done by applying *grid search* from 0.0 to 1.0 (step size: 0.01) using the validation set and F1-score macro as measure.
  - For English ensemble: 0.44; for non-English ensemble: 0.49.
- Pruned soft voting showed marginal improvement when applied to the validation set:

Voting	**F1-Score Macro (after removing 'no label')**
Non-Pruned	0.3807
Pruned	0.3902

Submission Results & Future Work

Subtask 1

Team	F1-Score (Macro)
Arthur Schopenhauer	0.35
Baseline	0.24

Subtask 2

Team	F1-Score (Macro)
Arthur Schopenhauer (Best Submission)	0.83
Baseline	0.81

Future Work

- Using larger model architectures.
- Fine-tuning different models to detect only certain values, rather than all 19 values.

Thank you!