ADU Classifier

using Deep Learning with Transformers

Telmo Baptista | M.EIC | FEUP

Recap

Our objective is to classify the propositional content of ADUs from the Público newspaper into:

- Facts;
- Policies;
- Negative Value;
- Neutral Value;
- Positive Value.

There's two datasets:

- Articles
  - Contains the raw article and its metadata;
- ADUs
  - Contains the ADUs for the articles;
  - ADU may be annotated multiple times by different annotators;
  - Has its corresponding label.

Base Language Model

- BERTimbau;
- BERT-based for Portuguese language;
- Trained on BrWaC (Brazilian Web as Corpus);
- High scores on multiple tasks compared to its peers;
- Contains base and large version:
  - Base - 110M parameters;
  - Large - 335M parameters (version used).

«BERTimbau Large is a pretrained BERT model for Brazilian Portuguese that achieves state-of-the-art performances on three downstream NLP tasks: Named Entity Recognition, Sentence Textual Similarity and Recognizing Textual Entailment»

Base Model in HuggingFace
https://huggingface.co/neuralmind/bert-large-portuguese-cased

Domain Adaptation

- Need to adapt language model to our domain:
  - Different language: European Portuguese vs Brazilian Portuguese;
  - Different context: News articles vs Variety of websites;
- Trained using MLM (Masked Language Modelling) → objective is to obtain a better representation of input documents;
- Intrinsic Evaluation for performance (perplexity).

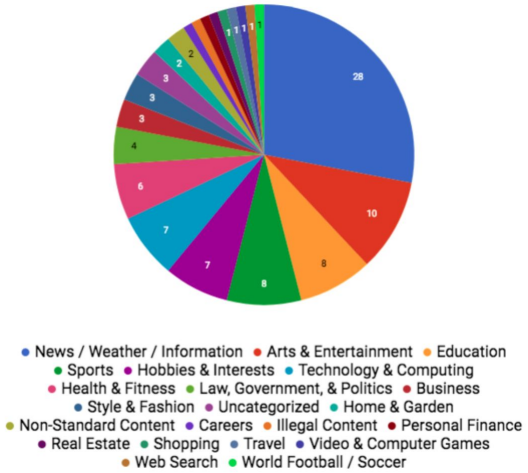

Fig. 1 - brWaC Corpus website content distribution. Filho et al., 2018.

Domain Adaptation

- Trained on articles dataset:
  - Training: 2914 chunks of 128 tokens (80%);
  - Validation: 729 chunks of 128 tokens (20%);
- Masking probability: 20%;
- Mini-batch training with size of 16 (shuffled);
- AdamW optimizer:
  - Initial learning rate: 1e-5;
  - Additional linear decay on each epoch;
- Trained for 10 epochs:
  - Best epoch: 10;
  - Perplexity: 5.257 (-44% compared to 1st epoch).

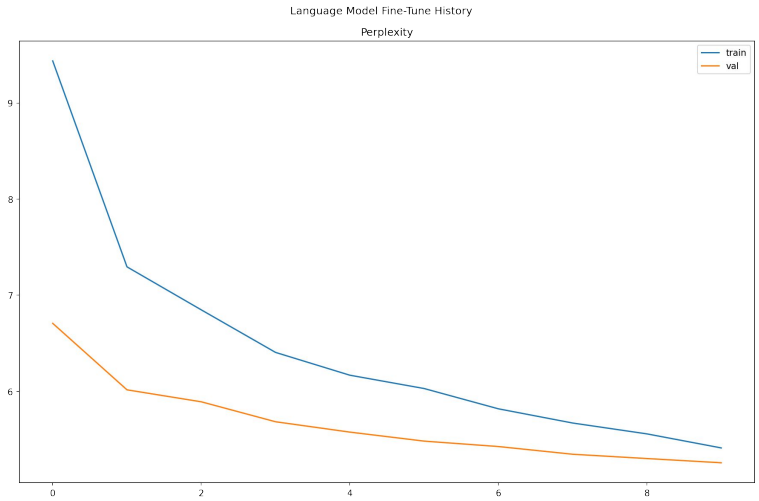

Fig. 2 - Perplexity evolution during MLM training.

Text Classification - *CLS_C*

- Trained on complete ADU dataset:
  - Training: 10716 ADUs (64%);
  - Validation: 2679 ADUs (16%);
  - Test: 3348 ADUs (20%)
  - Splits were made to ensure distribution of labels is maintained;
- Dynamic data augmentation:
  - Random replacement of words by a random synonym;
  - Replacement probability: 30%;
- Trained in two steps:
  - Freeze training;
  - Fine-tuning;
- Objective: minimize loss in the validation set;
- Evaluated on a variety of metrics:
  - Micro-accuracy, macro precision, recall, F1 score, others.

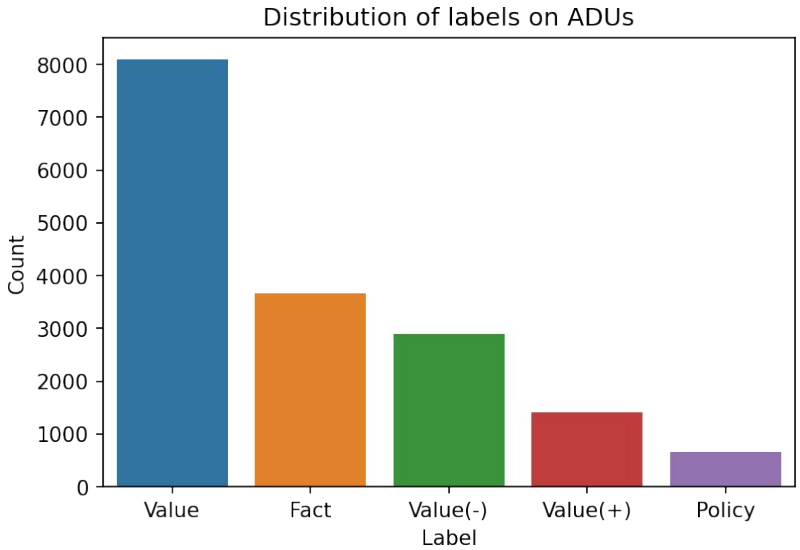

Fig. 3 - Label distribution on the ADUs dataset (complete dataset).

Text Classification - Freeze Training

- Language model layers are frozen;
- Mini-batch training with size of 64 (shuffled);
- Adam optimizer:
  - Initial learning rate: 5e-3;
  - Exponential decay on each epoch;
- Trained for 30 epochs:
  - Best epoch: 26;
  - Validation loss: 1.016;
  - Validation micro-accuracy: 58.83%;
- Small improvement to expected;
- Model stagnated → more epochs won't improve performance;

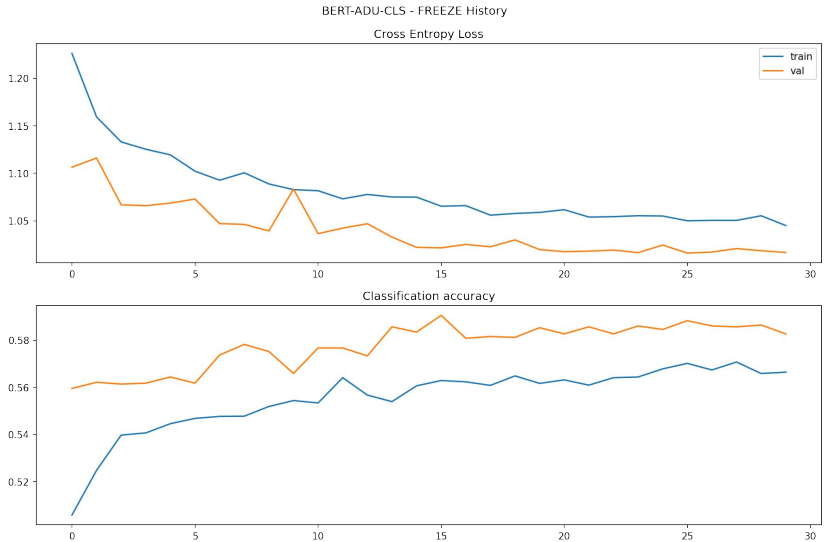

Fig. 4 - Loss and micro-accuracy evolution on freeze training stage.

Text Classification - Fine-tuning

- Language model layers are now unfrozen;
- Objective: try to push model out of local minima;
- Mini-batch training with size of 16 (shuffled);
- Adam optimizer:
  - Initial learning rate: 1e-5;
  - Exponential decay on each epoch;
- Trained for 15 epochs:
  - Best epoch: 1;
  - Validation loss: 0.876;
  - Validation micro-accuracy: 63.41%;
- Overfitted after first epoch.

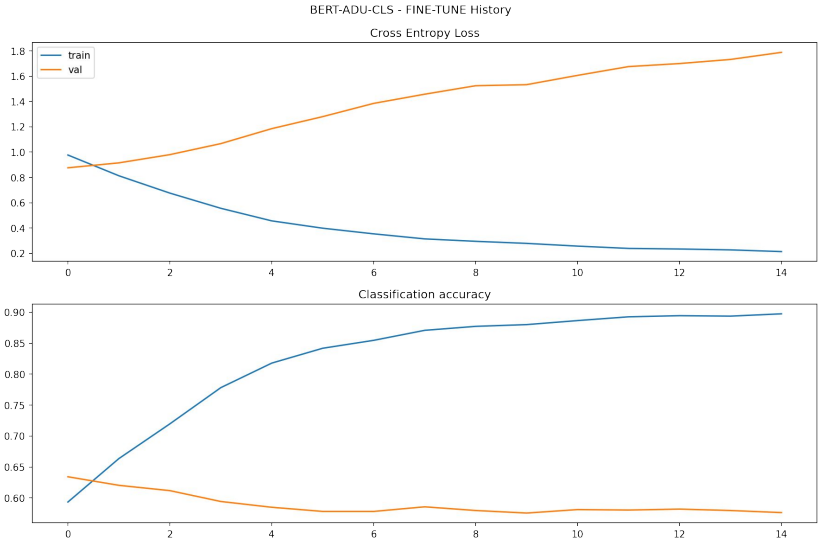

Fig. 5 - Loss and micro-accuracy evolution on fine-tuning stage.

Text Classification - Evaluation

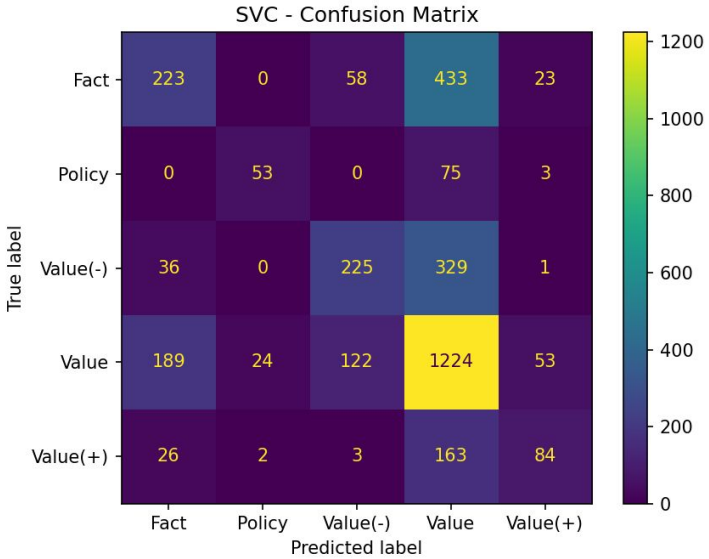

Fig. 6 - Confusion matrix for SVC (first project).

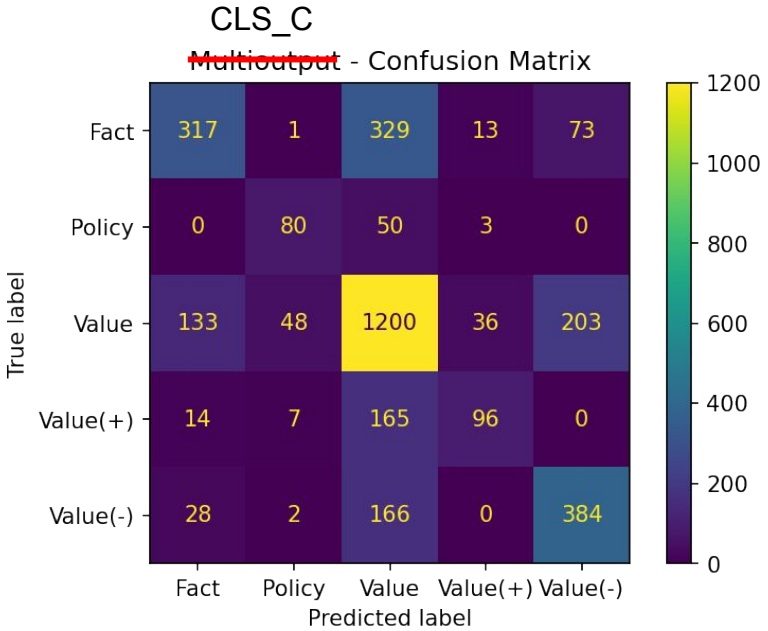

Fig. 7 - Confusion matrix for CLS_C.

Text Classification - Evaluation

SVC:

- Micro-accuracy: 54%

	Precision	Recall	F1-Score
Fact	0.47	0.30	0.36
Policy	0.67	0.40	0.50
Value	0.55	0.75	0.63
Value(+)	0.51	0.30	0.38
Value(-)	0.55	0.38	0.45
Weighted Avg	0.53	0.54	0.52

Table. 1 - Classification report for SVC (first project).

CLS_C:

- Micro-accuracy: 62%

	Precision	Recall	F1-Score
Fact	0.64	0.43	0.52
Policy	0.58	0.60	0.59
Value	0.63	0.74	0.68
Value(+)	0.65	0.34	0.45
Value(-)	0.58	0.66	0.62
Weighted Avg	0.62	0.62	0.61

Table. 2 - Classification report for CLS_C.

Text Classification - Evaluation

- Is the accuracy obtained really accurate? **No.**
- There's conflicts in annotations on the dataset, and they weren't treated here, let's look at some "errors".

ADU	Target	Predicted
"Em dezembro do ano passado Fernando Medina avançou com a proposta de obras profundas na Segunda Circular"	Value	Fact
"É tudo cómico na FIFA"	Value(+)	Value(-)
"Referindo-se aos incidentes, o presidente da UEFA, Michel Platini, afirmou estar muito triste com o sucedido"	Value(-)	Fact

Table. 3 - CLS_C errors on testing set.

Text Classification - *CLS_NC*

- Same in architecture to *CLS_C* but trained on dataset that doesn't contain conflicted ADUs:
  - Conflicts are resolved by majority wins, if tied both are eliminated;
- Trained on non-conflict ADU dataset:
  - Training: 6936 ADUs (64%);
  - Validation: 1469 ADUs (16%);
  - Test: 1787 ADUs (20%)
  - Splits were made to ensure distribution of labels is maintained;

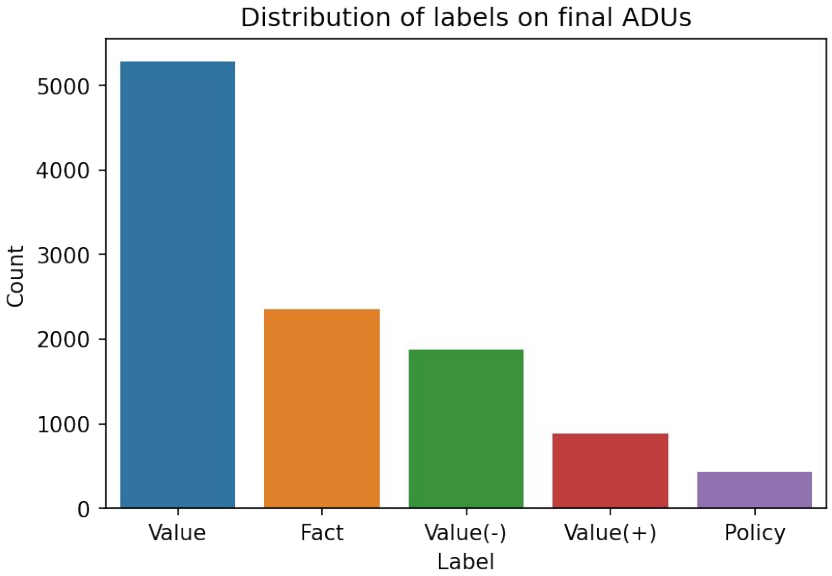

Fig. 8 - Label distribution after removing conflicted ADUs.

Text Classification - Training stages

- Freeze training for 30 epochs:
  - Best epoch: 25;
  - Validation loss: 1.011;
  - Validation micro-accuracy: 57.77%;

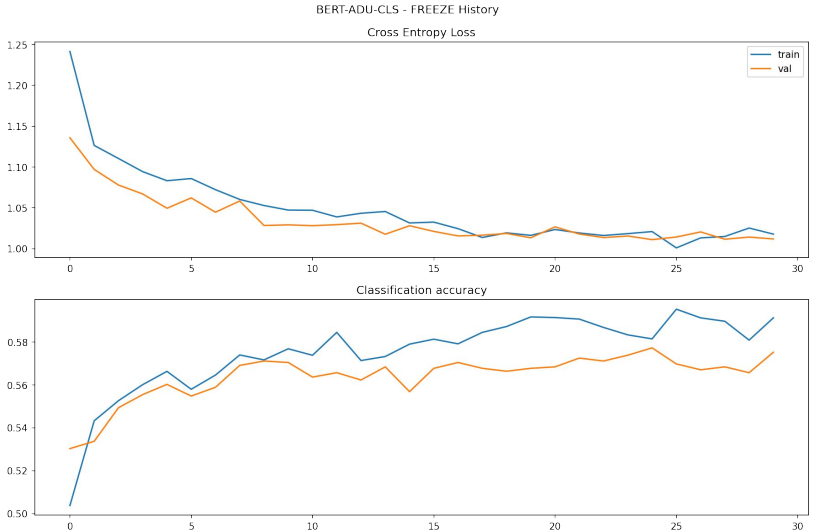

Fig. 9 - Loss and micro-accuracy evolution on freeze training stage.

- Fine-tuning for 15 epochs:
  - Best epoch: 1;
  - Validation loss: 0.869;
  - Validation micro-accuracy: 63.78%;

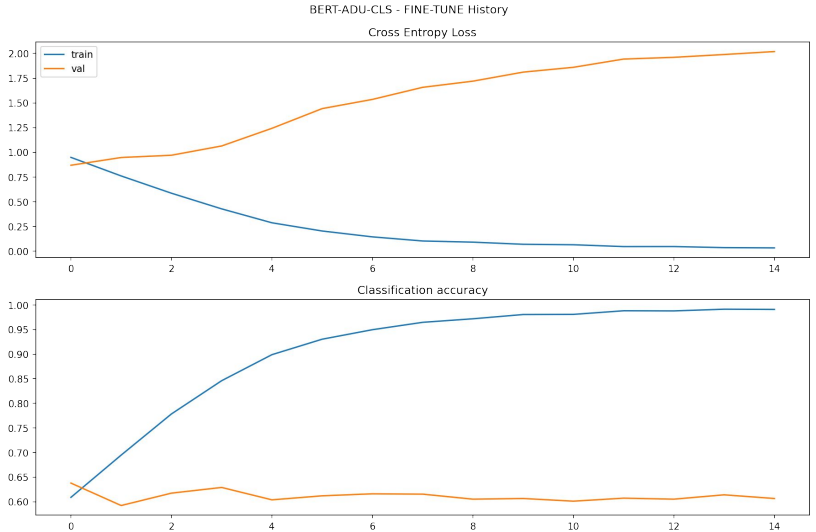

Fig. 10 - Loss and micro-accuracy evolution on fine-tuning stage.

Text Classification - Evaluation

CLS_NC:

- Micro-accuracy: 64%

	Precision	Recall	F1-Score
Fact	0.64	0.47	0.54
Policy	0.59	0.58	0.58
Value	0.67	0.74	0.70
Value(+)	0.59	0.41	0.49
Value(-)	0.59	0.69	0.63
Weighted Avg	0.64	0.64	0.63

Table. 4 - Classification report for CLS_NC.

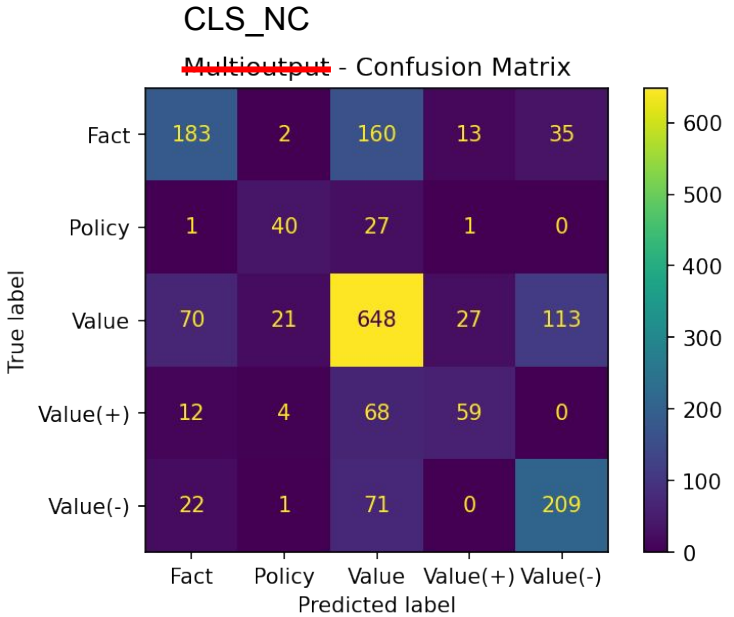

Fig. 11 - Confusion matrix for CLS_NC.

Mix-and-Match Experiment

What if we now use the classifier trained on conflicted dataset to classify the non-conflicted dataset?

CLS_NC:

- Micro-accuracy: 64%

	Precision	Recall	F1-Score
Fact	0.64	0.47	0.54
Policy	0.59	0.58	0.58
Value	0.67	0.74	0.70
Value(+)	0.59	0.41	0.49
Value(-)	0.59	0.69	0.63
Weighted Avg	0.64	0.64	0.63

Table. 4 - Classification report for CLS_NC.

CLS_C:

- Micro-accuracy: 70%

	Precision	Recall	F1-Score
Fact	0.71	0.48	0.57
Policy	0.68	0.75	0.72
Value	0.69	0.82	0.75
Value(+)	0.77	0.42	0.54
Value(-)	0.69	0.74	0.71
Weighted Avg	0.70	0.70	0.69

Table. 5 - Classification report for CLS_C in the non-conflict dataset.

Text Classification - Fine-tuning[2]

- Fine-tune *CLS_C* model on the dataset without conflicted ADUs;
- Language model frozen → unfrozen lead to overfit previously;
- Objective: try to push model out of local minima;
- Mini-batch training with size of 64 (shuffled);
- Adam optimizer:
  - Initial learning rate: 1e-4;
- Trained for 10 epochs:
  - Best epoch: 10;
  - Validation loss: 0.715;
  - Validation micro-accuracy: 70.45%;

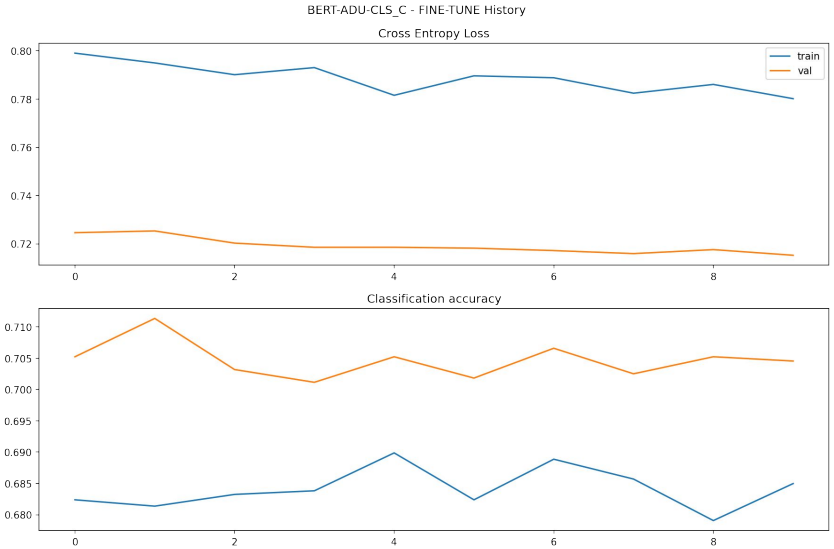

Fig. 12 - Loss and micro-accuracy evolution on fine-tuning stage.

Text Classification - Evaluation^2

CLS_C:

- Micro-accuracy: 70%

	Precision	Recall	F1-Score
Fact	0.71	0.48	0.57
Policy	0.68	0.75	0.72
Value	0.69	0.82	0.75
Value(+)	0.77	0.42	0.54
Value(-)	0.69	0.74	0.71
Weighted Avg	0.70	0.70	0.69

Table. 6 - Classification report for CLS_C in the non-conflict dataset.

CLS_C^2:

- Micro-accuracy: 71%

	Precision	Recall	F1-Score
Fact	0.70	0.52	0.60
Policy	0.65	0.78	0.71
Value	0.71	0.82	0.76
Value(+)	0.74	0.51	0.61
Value(-)	0.71	0.71	0.71
Weighted Avg	0.71	0.71	0.70

Table. 7 - Classification report for CLS_C^2 in the non-conflict dataset.

Wrapping it up

- Proposed model: **CLS_C²**;

	SVC*	CLS_NC	CLS_C	CLS_C²
Precision	53%	64%	70%	71%
Recall	54%	64%	70%	71%
F1-Score	52%	63%	69%	70%
Accuracy	54%	64%	70%	71%

Precision, recall and F1-scores are the weighted averages, and accuracy corresponds to the micro-accuracy on the non-conflicted ADU dataset.
*This model wasn't tested on the non-conflicted ADU dataset, but in the conflicted ADU dataset.

Table. 8 - Performance of developed models on the non-conflicted ADU dataset (see note above).

Conclusions & Future work

- The tuning of the language model in the domain adaptation phase wasn't deeply explored, so there's potential improvement there, including making data augmentation on that phase (which wasn't performed);
- The developed models all stagnated at the final results shown, one approach that wasn't implemented here but can potentially improve the performance is passing the neighbourhood of the ADU from the article as context for the model (*"Context matters!: identifying argumentative relations in essays"*, https://dl.acm.org/doi/10.1145/3477314.3507246);
- In the mix-n-match phase it was also tested how the model *CLS_NC* performed on the conflicted dataset, and it also was better than the original model trained there (achieving 66% compared to 62% by *CLS_C*), but the fine-tuning wasn't explored and can be another experiment. Although, it will most likely stagnate like *CLS_C*[2].
- Lastly, the quality of the dataset itself is also a factor on the performance of these models, and could be treated better to achieve better results.

Thanks

Any questions?