MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer

Ilias Chalkidis [†] [◇] **Manos Fergadiotis** [‡] **Ion Androutsopoulos** [‡]
[†] Department of Computer Science, University of Copenhagen, Denmark
[‡] Department of Informatics, Athens University of Economics and Business, Greece
[◇] Cognitiv+ Ltd., London, United Kingdom
ilias.chalkidis@di.ku.dk [fergadiotis,ion]@aueb.gr

Abstract

We introduce MULTI-EURLEX, a new multilingual dataset for topic classification of legal documents. The dataset comprises 65k European Union (EU) laws, officially translated in 23 languages, annotated with multiple labels from the EUROVOC taxonomy. We highlight the effect of temporal concept drift and the importance of chronological, instead of random splits. We use the dataset as a testbed for zero-shot cross-lingual transfer, where we exploit annotated training documents in one language (source) to classify documents in another language (target). We find that fine-tuning a multilingually pretrained model (XLM-ROBERTA, MT5) in a single source language leads to catastrophic forgetting of multilingual knowledge and, consequently, poor zero-shot transfer to other languages. Adaptation strategies, namely partial fine-tuning, adapters, BITFIT, LNFIT, originally proposed to accelerate fine-tuning for new end-tasks, help retain multilingual knowledge from pretraining, substantially improving zero-shot cross-lingual transfer, but their impact also depends on the pretrained model used and the size of the label set.

1 Introduction

Multilingual learning is an active field of research in NLP. Starting from neural machine translation (Stahlberg, 2020), multilingual neural models are increasingly being considered across NLP tasks and multilingual benchmark datasets for cross-lingual language understanding are becoming available (Hu et al., 2020; Ruder et al., 2021), complementing previous monolingual benchmarks (Wang et al., 2018). The initial paradigm of multilingual word embeddings (Ruder et al., 2017) was rapidly expanded to pretrained multilingual models (Conneau et al., 2018), including work on zero-shot cross-lingual transfer (Artetxe and Schwenk, 2019). Multilingual models based on TRANSFORMERS (Vaswani et al., 2017), jointly pretrained on large corpora across multiple languages, have significantly advanced the state-of-the-art in cross-lingual tasks (Conneau et al., 2020; Xue et al., 2021).

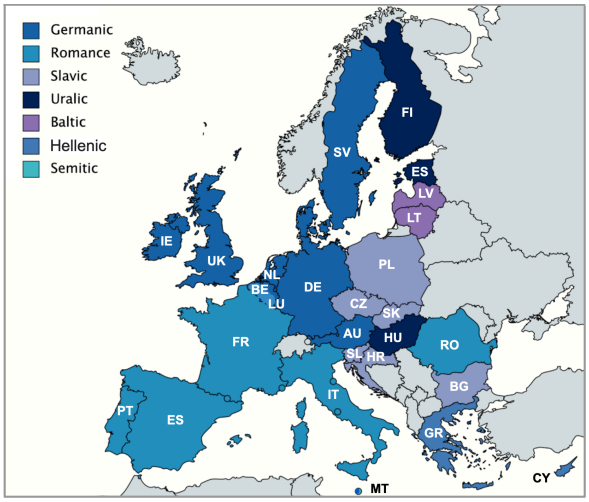

Figure 1: MULTI-EURLEX covers 23 official EU languages (Table 1) from 7 families (illustrated per EU country in the map). The UK was an EU member until 2020. The map should not be taken to imply that no other languages are spoken in EU countries.

In another interesting direction, legal NLP (Aletras et al., 2019; Zhong et al., 2020) is an emerging field targeting tasks such as legal judgment prediction (Aletras et al., 2016), legal topic classification (Chalkidis et al., 2019), legal question answering (Kim et al., 2015), contract understanding (Hendrycks et al., 2021), to name a few. Generic pretrained language models for legal text in particular have also been introduced (Chalkidis et al., 2020b). But despite rapid growth, cross-lingual transfer has not yet been explored in legal NLP.

To facilitate research on cross-lingual transfer for text classification and legal topic classification in particular, we introduce a new multilingual dataset, MULTI-EURLEX, which includes 65k European Union (EU) laws, officially translated in the 23 EU official languages (Fig. 1). Each document is annotated with multiple labels from EUROVOC, where concepts are organized hierarchi-

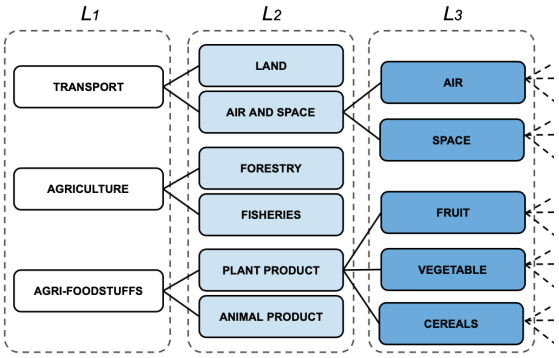

Figure 2: Examples from levels (L_i) 1 to 3 from the EUROVOC hierarchy. More general concepts become more specific as we move from higher to lower levels.

cally (Fig. 2).[1] We use the dataset as a testbed for zero-shot cross-lingual transfer in cases where we wish to exploit labeled training documents in one language (source) to classify documents in another language (target). This would allow, e.g., classifiers trained in resource-rich languages to be reused in languages with fewer or no training instances.

We experiment with monolingual and multilingual TRANSFORMER-based models, i.e., monolingual BERT models (Devlin et al., 2019), XLM-ROBERTA (Conneau et al., 2020), and MT5 (Xue et al., 2021). We find that fine-tuning a multilingual model in a single source language leads to catastrophic forgetting of multilingual knowledge and, consequently, poor zero-shot transfer to target languages. We show that adaptation strategies, namely, not fine-tuning some layers, adapters (Houlsby et al., 2019), BITFIT (Zaken et al., 2021), and LNFIT inspired by Frankle et al. (2021), originally proposed to accelerate fine-tuning for new end-tasks, help retain multilingual knowledge from pretraining, substantially improving zero-shot cross-lingual transfer, but their impact also depends on the particular pretrained model used and the size of the label set. We also compare chronological vs. random splits, highlighting the impact of temporal concept drift in legal topic classification, which causes random splits to over-estimate performance (Søgaard et al., 2021). Our main contributions are:

- A parallel multilingual annotated dataset for legal topic classification with 65k EU laws in 23 languages, which can be used as a testbed for cross-lingual multi-label classification.
- Extensive experiments with state-of-the-art monolingual and multilingual models in 23 languages, which establish strong baselines for research on cross-lingual (legal) text classification.
- Experiments with several adaptation strategies showing that adaptation is beneficial in zero-shot cross-lingual transfer, apart from task transfer.
- Comparison of chronological vs. random splits, showing the temporal concept drift in legal topic classification and problems with random splits.

[1] http://eurovoc.europa.eu/

2 Related Work

Legal topic classification has been studied for EU legislation (Mencia and Fürnkranz, 2007; Chalkidis et al., 2019) in a monolingual setting (English). While there are several legal NLP studies with non-English datasets (Kim et al., 2015; Waltl et al., 2017; Nguyen et al., 2018; Angelidis et al., 2018; Luz de Araujo et al., 2020), cross-lingual transfer has not been studied in the legal domain.

Cross-lingual transfer is a very active area of wider NLP research, currently dominated by large multilingually pretrained models (Conneau et al., 2018; Eisenschlos et al., 2019; Liu et al., 2020; Xue et al., 2021). Recent work explores adapter modules (Houlsby et al., 2019) to transfer monolingually pretrained (Artetxe et al., 2020) or multilingually pretrained (Pfeiffer et al., 2020) models to new (target) languages. We examine more adaptation strategies, apart from adapter modules, in truly zero-shot cross-lingual transfer. Unlike Pfeiffer et al. (2020), we do not train language-specific adapters per target language; we use adapters to fine-tune a *single* multilingual model on the source language, which is then used in all target languages.

In the broader field of multilingual legal studies, Gonalves and Quaresma (2010) examined legal topic classification with a dataset comprising 2.7k EU laws in 4 languages (English, German, Spanish, Portuguese). They experimented with monolingual SVM classifiers and their combination as a multilingual ensemble. More recently, Galassi et al. (2020) transferred sentence-level gold labels from annotated English to non-annotated German sentences, for the task of identifying unfair clauses in Terms of Service (2.7k sentences) and Privacy Policy documents (1.8k). They experimented with similarity-based methods aligning the English sentences to machine-translations of the German sentences. We experiment with state-of-the-art multilingual TRANSFORMER-based models considering many more languages (23) and a much larger dataset (65k EU laws). Although MULTI-EURLEX is largely parallel, we use it as a testbed for zero-shot cross-lingual transfer, *without* requiring parallel training data or machine translation systems.

Language	ISO code	Member Countries where official	EU Speakers (%)		Number of Documents			Words per document
			Native	Total	Train	Dev.	Test	
English	**en**	United Kingdom (1973–2020), Ireland (1973), Malta (2004)	13%	51%	55,000	5,000	5,000	1200 / 460
German	**de**	Germany (1958), Belgium (1958), Luxembourg (1958)	16%	32%	55,000	5,000	5,000	1085 / 410
French	**fr**	France (1958), Belgium(1958), Luxembourg (1958)	12%	26%	55,000	5,000	5,000	1280 / 480
Italian	**it**	Italy (1958)	13%	16%	55,000	5,000	5,000	1210 / 460
Spanish	**es**	Spain (1986)	8%	15%	52,785	5,000	5,000	1380 / 530
Polish	**pl**	Poland (2004)	8%	9%	23,197	5,000	5,000	1200 / 420
Romanian	**ro**	Romania (2007)	5%	5%	15,921	5,000	5,000	1500 / 500
Dutch	**nl**	Netherlands (1958), Belgium (1958)	4%	5%	55,000	5,000	5,000	1230 / 470
Greek	**el**	Greece (1981), Cyprus (2008)	3%	4%	55,000	5,000	5,000	1230 / 470
Hungarian	**hu**	Hungary (2004)	3%	3%	22,664	5,000	5,000	1120 / 370
Portuguese	**pt**	Portugal (1986)	2%	3%	23,188	5,000	5,000	1290 / 500
Czech	**cs**	Czech Republic (2004)	2%	3%	23,187	5,000	5,000	1170 / 410
Swedish	**sv**	Sweden (1995)	2%	3%	42,490	5,000	5,000	1130 / 470
Bulgarian	**bg**	Bulgaria (2007)	2%	2%	15,986	5,000	5,000	1480 / 510
Danish	**da**	Denmark (1973)	1%	1%	55,000	5,000	5,000	1080 / 410
Finnish	**fi**	Finland (1995)	1%	1%	42,497	5,000	5,000	890 / 320
Slovak	**sk**	Slovakia (2004)	1%	1%	15,986	5,000	5,000	1180 / 410
Lithuanian	**lt**	Lithuania (2004)	1%	1%	23,188	5,000	5,000	1070 / 370
Croatian	**hr**	Croatia (2013)	1%	1%	7,944	2,500	5,000	1490 / 500
Slovene	**sl**	Slovenia (2004)	<1%	<1%	23,184	5,000	5,000	1170 / 400
Estonian	**et**	Estonia (2004)	<1%	<1%	23,126	5,000	5,000	950 / 330
Latvian	**lv**	Latvia (2004)	<1%	<1%	23,188	5,000	5,000	1080 / 380
Maltese	**mt**	Malta (2004)	<1%	<1%	17,521	5,000	5,000	1250 / 430

Table 1: MULTI-EURLEX statistics per language: ISO code; EU countries using the language officially (year the country joined the EU in brackets); percentage of EU population speaking the language natively or in total (as native or non-native speakers);[3] documents in training, development, test splits; words per document (mean/median).

3 The MULTI-EURLEX Dataset [2]

Documents: MULTI-EURLEX comprises 65k EU laws in 23 official EU languages (Table 1). Each EU law has been annotated with EUROVOC concepts (labels) by the Publications Office of EU. Each EUROVOC label ID is associated with a *label descriptor*, e.g., ⟨60, 'agri-foodstuffs'⟩, ⟨6006, 'plant product'⟩, ⟨1115, 'fruit'⟩. The descriptors are also available in the 23 languages. Chalkidis et al. (2019) published a *monolingual* (English) version of this dataset, called EURLEX57K, comprising 57k EU laws with the originally assigned gold labels.

Languages: MULTI-EURLEX covers 23 languages from 7 families (Fig. 1). EU laws are published in all official EU languages, except for Irish for resource-related reasons.[3] This wide coverage makes the dataset a valuable testbed for cross-lingual transfer. All languages use the Latin script, except for Bulgarian (Cyrillic script) and Greek.

Label Set	No. of Labels	In training docs		In all docs	
Level 1	21	21	(100%)	21	(100%)
Level 2	127	127	(100%)	127	(100%)
Level 3	567	500	(88%)	511	(90%)
All	7,390	4,220	(57%)	4,591	(62%)

Table 2: EUROVOC concepts in the four label sets and how many are used in the training or entire dataset.

Multi-granular Labeling: EUROVOC has eight levels of concepts (Fig. 2 illustrates three). Each document is assigned one or more concepts (labels). If a document is assigned a concept, the ancestors and descendants of that concept are typically not assigned to the same document. The documents were originally annotated with concepts from levels 3 to 8. We created three alternative sets of labels per document, by replacing each assigned concept by its ancestor from level 1, 2, or 3, respectively. Thus, we provide four sets of gold labels per document, one for each of the first three levels of the hierarchy, plus the original sparse label assignment.[5] Table 2 presents the distribution of labels across label sets.

Supported Tasks: Similarly to EURLEX57K (Chalkidis et al., 2019), MULTI-EURLEX can be used for legal topic classification, a multi-label classification task where legal documents need to

[2] The dataset is available at https://huggingface.co/datasets/multi_eurlex. Following Gebru et al. (2018), we provide an extended Dataset Card in Appendix D.

[3] https://europa.eu/european-union/about-eu/eu-languages_en

[4] Data from European Commision (2012). Following BREXIT (2020), UK citizens are no longer considered EU citizens, thus native English speakers became approx. 1% in the EU, as of 2021. Table 1 includes UK citizens.

[5] Levels 4 to 8 cannot be used independently, as many documents have gold concepts from the third level; thus many documents will be mislabeled, if we discard level 3.

be assigned concepts (in our case, from EUROVOC) reflecting their topics. Unlike EURLEX57K, however, MULTI-EURLEX supports labels from three different granularities (EUROVOC levels). More importantly, apart from monolingual (*one-to-one*) experiments, it can be used to study cross-lingual transfer scenarios, including *one-to-many* (systems trained in one language and used in other languages with no training data), and *many-to-one* or *many-to-many* (systems jointly trained in multiple languages and used in one or more other languages).

Data Split and Concept Drift: MULTI-EURLEX is *chronologically* split in training (55k, 1958–2010), development (5k, 2010–2012), test (5k, 2012–2016) subsets, using the English documents. The test subset contains the same 5k documents in all 23 languages (Table 1).[6] For the official languages of the seven oldest member countries, the same 55k training documents are available; for the other languages, only a subset of the 55k training documents is available (Table 1). Compared to EURLEX57K (Chalkidis et al., 2019), MULTI-EURLEX is not only larger (8k more documents) and multilingual; it is also more challenging, as the chronological split leads to temporal real-world *concept drift* across the training, development, test subsets, i.e., differences in label distribution and phrasing, representing a realistic *temporal generalization* problem (Huang and Paul, 2019; Lazaridou et al., 2021). Recently, Søgaard et al. (2021) showed this setup is more realistic, as it does not over-estimate real performance, contrary to random splits (Gorman and Bedrick, 2019).

Label Set	Random		Chronological	
	train-dev	*train-test*	*train-dev*	*train-test*
Level 1	0.00	0.00	0.03	0.04
Level 2	0.00	0.00	0.12	0.16
Level 3	0.01	0.01	0.21	0.32
All	0.20	0.20	1.09	1.67

Table 3: KL-divergence of label distributions between subsets, using a *random* or *chronological* split.

To verify that the chronological split of MULTI-EURLEX in training, development, test subsets leads to a *temporal concept drift*, we compare the KL-divergence between the label distributions of the subsets using the chronological vs. a random split. Table 3 shows a random split leads to almost zero divergence for levels 1–3 and low divergence when using all labels. With the chronological split, the divergence increases as the number of labels increases, and is larger between the train and test subsets, which have a larger temporal distance compared to the train and development subsets.

Data Split	Training	Development	Test
Random	**99.2**	**74.7**	**74.0**
Chronological	96.7	58.7	48.4

Table 4: Results of MULTI-EURLEX for the original sparse annotation (7,390 labels) with BERT using a *random* or *chronological* split. Here the model is fine-tuned and tested on English data only (*one-to-one*).

To further highlight the temporal concept drift, we fine-tune BERT (Devlin et al., 2019) on the English part of MULTI-EURLEX using all labels, following Chalkidis et al. (2019). Table 4 shows that although the performance on training data is very high with both splits, it deteriorates more rapidly on development data with the chronological split. Also, performance is stable when moving from development to test data with the random split, since both subsets contain randomly sampled unseen documents; but with the chronological split, performance continues to decline on test data. This confirms our hypothesis of a temporal concept drift and shows that the random split over-estimates real performance, contrary to the chronological split.

4 Methods

4.1 Pretrained Models

NATIVE-BERTS: Many monolingual pretrained TRANSFORMER-based (Vaswani et al., 2017) models have been released, based on BERT (Devlin et al., 2019) or ROBERTA (Liu et al., 2019).[7] Across classification experiments, L is the cardinality of the label set, and D_h the dimensionality of the hidden states. We feed the top-level hidden state of the `[cls]` token ($\in \mathbb{R}^{D_h}$) to a dense layer ($W_{\texttt{[cls]}} \in \mathbb{R}^{D_h \times L}$) with L outputs and sigmoids.

XLM-ROBERTA: Conneau et al. (2020) introduced a multilingual ROBERTA for 100 languages. It is pretrained on Common Crawl with a vocabulary of 250k sub-words shared across languages. We use the same classification setup as in NATIVE-BERTS.

MT5: Xue et al. (2021) released a multilingual variant of T5 (Raffel et al., 2020), an encoder-decoder TRANSFORMER-based model, pretrained on text in 101 languages from Common Crawl. As in T5, Xue et al. (2021) frame all NLP tasks (incl. text

[6] The development subset also contains the same 5k documents in 23 languages, except Croatian. Croatia is the most recent EU member (2013); older laws are gradually translated.

[7] Appendix C lists the native pretrained BERTs we used.

classification) as text generation. This approach (text-to-text) is reasonable in single-label multi-class classification tasks like those of GLUE (Wang et al., 2018), where the output is expected to be the textual descriptor of a single class. But in our case, we have a *multi-label* task with 5 labels per document on average and label sets containing hundreds or thousands of labels; hence a textual output would be unnecessarily complex. Also, requiring a *sequence* of labels as output would be problematic, since the correct labels are not ordered. Hence, we use only the encoder of MT5. Similarly to XLM-ROBERTA, we add a `[cls]` special token, always at the beginning of the sequence, and use its top-level hidden state to represent the document.[8]

4.2 Cross-lingual Adaptation Strategies

We mainly study *zero-shot cross-lingual transfer*, where we fine-tune (further train) a multilingual model (pretrained on a multilingual corpus) only on annotated documents of a *source* language, and evaluate it (without any further training) on test (and development) documents in the other 22 languages (*one-to-many*). To avoid *catastrophically forgetting* the multilingual pretraining when fine-tuning only for the source language, we examine adaptation strategies, where the model is only partially fine-tuned. These were originally proposed to accelerate fine-tuning when moving to new end-tasks, but we employ them to retain multilingual knowledge. The four strategies are the following:

Frozen layers: In this case, we follow Rosenfeld and Tsotsos (2019) and do not update the parameters of the first N or all (N =12) stacked TRANSFORMER blocks in fine-tuning; we also never update any input embeddings (of tokens, positions, segments). We experiment with $N = 3, 6, 9, 12$.

Adapter modules: In this case, we follow Houlsby et al. (2019), placing adapter modules after each feed-forward layer (FFNN) inside each TRANSFORMER encoder block. Each block contains two FFNN layers: one after the attention layer and one at the very end. An adapter module consists of a down-projection dense layer ($W_{down} \in \mathbb{R}^{D_h \times K}$, assuming row-vectors, where $K \ll D_h$) and a consecutive up-projection ($W_{up} \in \mathbb{R}^{K \times D_h}$), followed by a residual connection (He et al., 2016). The rest of the Transformer block is not updated, except for layer normalization components (Ba et al., 2016).

BitFit: BITFIT (Zaken et al., 2021) keeps the whole network frozen during fine-tuning, except for bias terms. Zaken et al. showed that applying BITFIT on the English BERT (updating 0.09% of parameters) is competitive with fully fine-tuning the entire model in the GLUE benchmark (Wang et al., 2018).

LNFit: Similarly, Frankle et al. (2021) train only the parameters of *batch normalization* (Ioffe and Szegedy, 2015) layers in image classifiers. We adopt a similar approach, dubbed LNFIT, where we fine-tune only the *layer normalization* parameters of pre-trained TRANSFORMERs for text.

The randomly-initialized classification (dense) layer on top of the encoder is always fine-tuned.

5 Experimental Setup

Configuration of Models and Training Details: We implemented all methods in TENSORFLOW 2, obtaining pretrained models from the Hugging Face library. We release our code and data for reproducibility.[9] All models follow the BASE configuration with 12 stacked TRANSFORMER encoder blocks, each with $D_h = 768$ and 12 attention heads. We use the Adam optimizer (Kingma and Ba, 2015) across all experiments. We grid-search to tune the learning rate per method, considering classification performance on development data.[10]

Evaluation: Given the large number and skewed distribution of labels, retrieval measures have been favored in large-scale multi-label text classification literature (Mullenbach et al., 2018; Chalkidis et al., 2019). Following Chalkidis et al. (2019, 2020a), we report *mean R-Precision* (MRP) (Manning et al., 2009). That is, for each document, the model ranks the labels it selects by decreasing confidence, and we compute Precision@k, where k is the document's number of gold labels; we then average over documents. For all experiments, we use the chronological data split and report the average across three runs. Unless stated otherwise, we use level 3 with $L = 567$ labels (Table 2), which has a highly skewed (long-tail) label distribution and temporal concept drift (Table 3). In Section 6.2, we also consider label sets from the other levels.

[8] In additional experiments, we also examined the original generative fine-tuning of MT5 (Xue et al., 2021), and another simplified encoder-decoder variant of MT5 agnostic of label order. Both led to worse performance, while being substantially larger (40% more parameters). See Appendix B.

[9] Our code is available on Github (`https://github.com/nlpaueb/multi-eurlex`).

[10] See Appendix A for details on hyper-parameter tuning.

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

Table 5: Test results for level 3 (567 labels) of MULTI-EURLEX. We show mRP (%) for the 16 most widely spoken EU official languages, and mRP averaged over all 23 languages. Appendix E reports results for all languages.

6 Experiments and Discussion

For the main experiments, we mainly use XLM-ROBERTA in a *one-to-many* setting (fine-tuning in English, testing in all languages). We also report key MT5 results for completeness. As a ceiling for cross-lingual transfer, in Section 6.1 we first evaluate monolingual (native) BERT models and XLM-ROBERTA, both in a *one-to-one* manner (fine-tuning and testing in the same language), which requires annotated training data in the target language. For completeness, in Section 6.3, we also report *many-to-many* results, where XLM-ROBERTA is jointly fine-tuned and tested in all languages.

6.1 Monolingual Classification (*one-to-one*)

Table 5 (top) shows that in the *one-to-one* setting, XLM-ROBERTA is competitive to native (monolingually pretrained) BERTs with a minor decrease of 0.7 mRP on average across languages. Of course, the *one-to-one* setting requires training data in the target language. We report these results as an *upper bound* for zero-shot cross-lingual transfer. Also, the native BERTs are pretrained on corpora of different sizes and quality, which explains why they are not consistently better than XLM-ROBERTA.

6.2 Cross-lingual Transfer (*one-to-many*)

XLM-ROBERTA adaptation: In the one-to-many setting, where we fine-tune in English and test in all languages, Table 5 (middle) shows that all adaptation strategies vastly improve the performance of XLM-ROBERTA across languages (up to 6.8 All mRP increase) comparing to no adaptation (end-to-end fine-tuning), while remaining competitive in English (source). This indicates that not fine-tuning the full set of parameters helps the model retain more of its multilingual knowledge obtained during pretraining. We observe no big difference among the block freezing strategies for $N = 3$, 6, 9, but performance deteriorates substantially when all blocks are frozen ($N = 12$).[11] We speculate there is a trade-off between freezing more blocks to retain multilingual knowledge and freezing fewer blocks to benefit end-task (classification) performance. Adapters consistently lead to the best results in all languages and overall (All mRP 56.1), with practically no decrease in English (source) performance (67.3). BITFIT, which only fine-tunes bias terms (4e-2% of parameters), and LNFIT, which fine-tunes even fewer parameters (1e-2%), are the second and third best strategies. These results highlight the expressive power of the few parameters BITFIT and LNFIT modify; this observation has been also discussed in previous studies (Frankle et al., 2021; Zaken et al., 2021), but not in a multi-lingual setting. Overall, fine-tuning in a single language leads to substantial forgetting of multilingual knowledge, but adaptation strategies, especially adapter modules, alleviate this problem and improve cross-lingual end-task performance.

MT5 adaptation: In Table 6, we repeat the *one-to-many* experiments of Table 5, this time with MT5 (encoder only). For brevity, we report only mRP on the source (English) language, and mRP averaged over the 23 languages.[12] BITFIT cannot be applied in this case, because MT5 does not use bias terms. As in Table 5, freezing the initial N blocks of the encoder ($N = 3, 6, 9$) improves cross-lingual transfer (average mRP increase up to 4.7), but freezing

[11] When $N = 12$, we practically evaluate (probe) the intact pre-training knowledge of XLM-ROBERTA in the end-task.

[12] See Appendix E for additional experimental results.

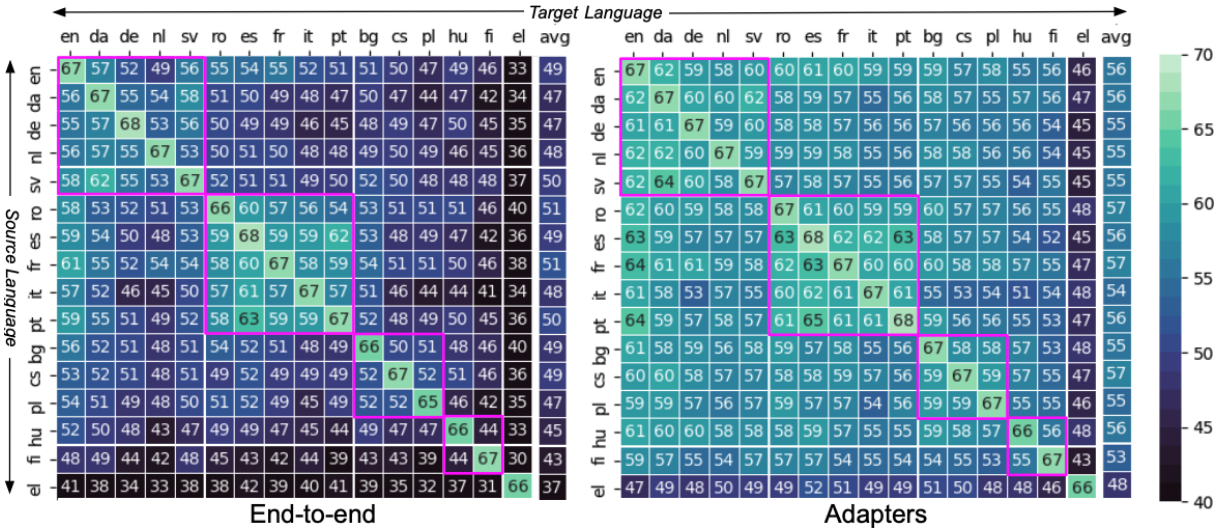

Figure 3: Test results (mRP, %) for Level 3 (567 labels) with XLM-ROBERTA, when fine-tuning in one language (source, rows) and testing in all languages (columns), without adaptation (end-to-end, left) and with adapter modules (right). The languages are grouped (framed) in language families (Germanic, Romance, Slavic, Uralic).

Adaptation strategy	Params (%)		en (Src)	All
End-to-end fine-tuning	277M	(100.0%)	67.4	53.7
First 3 blocks frozen	63.7M	(23.0%)	67.4	56.9
First 6 blocks frozen	42.4M	(15.3%)	66.3	**58.4**
First 9 blocks frozen	21.2M	(7.7%)	**68.0**	58.3
All 12 blocks frozen	–	(0.0%)	20.2	16.8
Adapter modules	7.1M	(1.7%)	66.3	44.0
LNFIT (layer-norm only)	19.2K	(0.01%)	59.5	38.7

Table 6: Test results of MT5 fine-tuned in English (en). We show mRP (%) in English (Src), and averaged across all 23 languages (All). We also report the trainable parameters (excl. the classification layer).

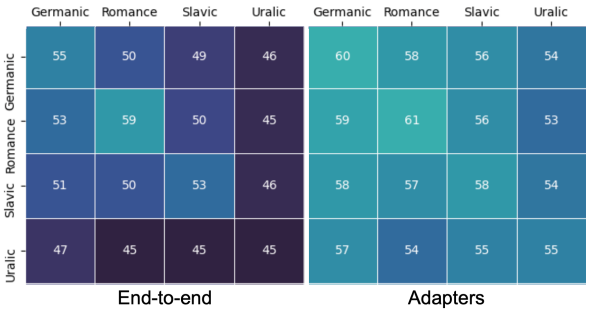

Figure 4: Cross-lingual test results (mRP, %) for level 3 (567 labels) with XLM-ROBERTA, averaged over language families (transfer from one family to another).

all layers ($N = 12$) harms performance. Surprisingly adapter modules, which are the best adaptation strategy for XLM-ROBERTA (Table 5, middle), lead to very poor performance (average mRP 44); there are similar results with LNFIT (average mRP 38.7). We speculate this happens because the encoder of MT5 needs to 're-program' itself during fine-tuning to perform as a stand-alone encoder; in adapters and LNFIT 're-programming' is only facilitated by very few parameters and the model is 'forced' (due to low adaptable capacity) to discard multilingual knowledge aggressively. XLM-ROBERTA follows the opposite pattern (fewer parameters lead to better cross-lingual transfer), because it is pre-trained as a stand-alone encoder. We leave a more thorough investigation of the trade-off between the number of trainable parameters vs. end-task (Src/All) performance for future work.

Different source languages: In the cross-lingual experiments so far, we fine-tuned the model in English (source) and evaluated it in all 23 languages. In Fig. 3, we repeat these experiments using a *different source language in each repetition* (rows), evaluating again in all languages (columns).[13] We use XLM-ROBERTA without adaptation (end-to-end, left) or with adapter modules (right). Despite the dominance of English in multi-lingual NLP literature, we observe that using alternative source languages (e.g., Romanian or French) lead to better target results. Similar results have been presented in Turc et al. (2021) for other NLP tasks. As in the previous one-to-many experiments with XLM-ROBERTA (Table 5, middle), adapters vastly improve cross-lingual transfer across all cases (e.g., English-en to Danish-da improves from 57 to 62 mRP), with occasionally slightly lower monolingual performance (e.g., German-de drops from 68 to 67). Cross-lingual transfer performs overall better when the source and target languages are in the same family (frames of Fig. 3), especially for Romance languages (Fig. 4, diagonal).[13] Also, when using adapters, cross-lingual performance often

[13] See Appendix E for additional experimental results.

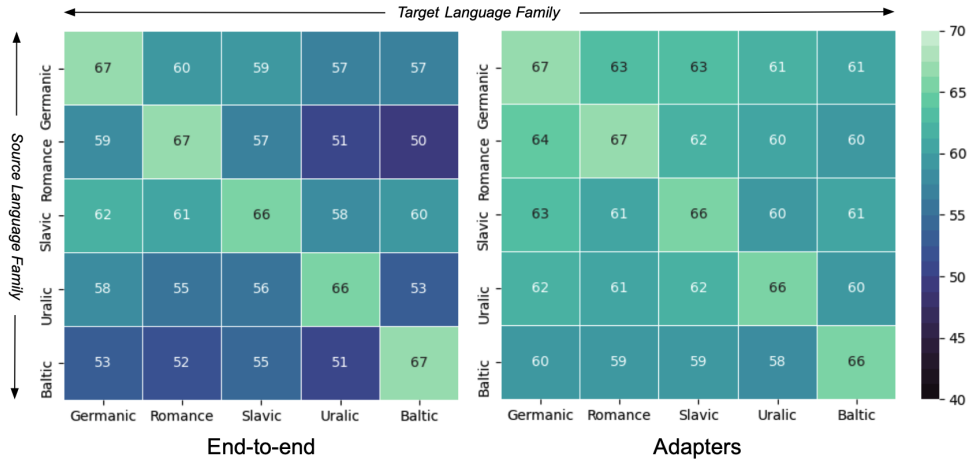

Figure 5: Cross-lingual test results (mRP, %) for level 3 (567 labels) with XLM-ROBERTA, when fine-tuning *end-to-end* or with adapters in *all languages of the same family (Src)* and testing is averaged over each language family.

drops less abruptly when moving outside of the family of the source language. For example, when fine-tuning in Danish-da, if the test set changes from Swedish-sv to Spanish-es, performance drops from 58 to 50 without adapters (Fig. 3, left), but the change is smoother, from 62 to 59, with adapters (Fig. 3, right). This is better illustrated in the right part of Fig. 4 (smoother changes across cells per row). These results confirm that adapter modules help retain more multilingual knowledge.

Transfer from one family to another: In the previous experiment ("*Different source languages*"), we used a *different source language in each repetition* and evaluated in all languages. To better understand how linguistic proximity between families affects performance, in Figure 5 we present additional experiments in a *many-to-many* setting, where each model is trained across *all languages in the same family* (source) and evaluated across all languages. We use again XLM-ROBERTA without adaptation (end-to-end, left) or with adapter modules (right). We observe (Fig. 5, left) that cross-lingual transfer performs overall better when the source and target families are the same. Also, when using adapters (Fig. 5, right), cross-lingual performance drops less abruptly when moving to another family, different from the one (source) whose languages were used for fine-tuning. As expected, the cross-lingual performance of these models (jointly fine-tuned in a language family) is substantially higher than the ones trained in a *one-to-one* setting (Fig. 3–4), and closer to that of the models jointly fine-tuned in *all 23 languages* (*many-to-many*, results reported in the lower part of Table 5).

Version of the input text	**en** (Src)		**Rest**	
	T	mRP	T (%)	mRP
Full-text	100%	**67.3**	100%	**56.1**
w/o digits	89%	67.1	88%	55.0
w/o digits & English vocab.	22%	14.0	77%	51.5

Table 7: Test results of XLM-ROBERTA (with adapters) removing digits and words used in the English part of MULTI-EURLEX during inference. We show mRP (%) for English (Src) and averaged over the other 22 languages (Rest). T is the percentage of tokens retained.

Removing digits and shared words: In an ablation study, during inference we remove digits and words that are shared across languages to see to what extent label predictions depend on them. Initially, we eliminate digits, which constitute approx. 10% of the average document length measured in white-space separated tokens. Digits often participate in legal references (e.g., *"established by Regulation No 1468/81"*) or other coding schemes that may hint EUROVOC concepts (e.g., when specific laws are highly cited). Moreover, inspecting training documents, we observe that vocabulary words (e.g., of Latin origin) are shared to a substantial degree (23% on average) across languages; thus as a second step we remove approx. 25k words used more than 25 times in English documents to break direct cross-lingual alignment. Table 7 shows that removing digits leads to a small decrease in *one-to-one* performance (-0.2) and a larger, though still small, decrease in *one-to-many* performance (-1.1). Eliminating shared words (present in the English vocabulary) leads to a further decrease (-3.5) in cross-lingual performance, and English performance of course plunges as the remaining text is very short and severely corrupted.

Adaptation Strategy	Parameters		Level 1 (21)		Level 2 (127)		Level 3 (567)		Original (7,390)	
			en (Src)	**All**	**en** (Src)	**All**	**en** (Src)	**All**	**en** (Src)	**All**
End-to-end fine-tuning	278M	(100%)	**83.2**	75.7	**73.6**	58.7	**67.4**	49.3	47.6	27.6
First 3 blocks frozen	63.8M	(23.0%)	82.9	76.4	71.3	60.2	66.3	53.0	47.3	29.0
First 6 blocks frozen	42.5M	(15.3%)	82.3	76.7	69.6	61.1	66.3	51.7	47.1	30.1
First 9 blocks frozen	21.3M	(7.7%)	82.0	74.8	70.7	60.1	65.8	53.0	48.0	32.8
Adapter modules	9.5M	(3.3%)	83.1	**77.2**	72.3	**61.2**	67.3	**56.1**	47.9	**35.1**
BITFIT (bias terms only)	101K	(0.04%)	82.7	76.1	70.2	60.1	63.9	53.7	**48.3**	33.9
LNFIT (layer-norm only)	36.8K	(0.01%)	81.5	74.9	69.7	59.3	63.1	53.3	43.1	26.4
↑ ***Averaged Adapt.*** ↑	-		82.4	76.0	70.6	60.3	65.5	53.5	47.0	31.2
All 12 blocks frozen	-	(0.0%)	61.4	56.5	39.0	31.6	27.2	22.2	26.1	15.3

Table 8: Test results of XLM-ROBERTA fine-tuned in English, for all adaptation strategies and different label granularities (EUROVOC levels, Table 2). We show mRP results (%) for English (Src) and averaged over all 23 languages (All). We also count the trainable parameters, excl. the classification layer, which remains the same.

Different label granularities: Table 8 shows XLM-ROBERTA results with labels from different EUROVOC levels (Table 2) for all adaptation strategies. As expected, performance deteriorates (approx. 5-10% per level) as the size of the label set increases. Nonetheless, we observe consistent improvements with adaptation strategies compared to full (end-to-end) fine-tuning for all label sets, with the exception of the fully (all 12 blocks) frozen model (last row). Adapters have the best overall performance, but the ranking and impact of the different adaptation strategies varies across levels. Specifically, as the size of the label set increases, the average (Table 8, second-to-last line) adaptation zero-shot (All) performance: (a) improves compared to no adaptation (end-to-end fine-tuning), approx. +0.3 → +1.6 → +4.2 → +3.6, as we move from level 1 to the full (original) label set, with a small drop from level 3 to the full label set; and (b) deteriorates more aggressively when comparing it to English (Src) performance, approx. -6.4 → -10.3 → -12.0 → -15.8. The latter (b) is due to the need to model increasingly finer concepts (labels), which complicates cross-lingual concept alignment and, hence, hurts transfer, leaving more scope for adaptation strategies to make a difference (a).

6.3 Multilingual Fine-tuning (*many-to-many*)

In the lower part of Table 5, we report results for XLM-ROBERTA, fine-tuned end-to-end or using adapters, when the model is *jointly fine-tuned in all languages*. In this case, for each epoch and batch we randomly select a language of the document, among the available ones; not all documents are available in all 23 languages (Table 1). Adapter modules again consistently improve performance. The *many-to-many* models largely outperform the *one-to-many* models (Table 5, middle), as they have access to annotated training documents in all languages. Nevertheless, this is still an interesting scenario, because it allows *deploying a single model* that handles all languages and is competitive to using multiple native BERT models, one per language

7 Conclusions and Future Work

We introduced MULTI-EURLEX, a new multilingual legal topic classification dataset with 65k documents (EU laws) in 23 languages, where each document is annotated with multiple labels (concepts) from the EUROVOC taxonomy, with alternative label granularities. To the best of our knowledge, this is one of the most diverse, in terms of languages, classification datasets. We mainly used the dataset as a testbed for zero-shot cross-lingual transfer.

Experimental results showed that fine-tuning a multilingually pretrained model (XLM-ROBERTA, MT5) in a single language leads to catastrophic forgetting of multilingual knowledge and, consequently, poor zero-shot transfer. We found that adaptation strategies, originally proposed to accelerate fine-tuning for end-tasks, help retain multilingual knowledge from pretraining, substantially improving zero-shot cross-lingual transfer. However, their impact depends on the size of the label set, i.e., the gains increase as the label set increases. Interestingly, even adaptation strategies (BITFIT, LNFIT) that fine-tune a very small fraction of parameters (<0.05%) are competitive. Experimental results also showed that multilingual models are competitive to monolingual models in the one-to-one set-up; and that a single multilingual model jointly fine-tuned in all languages is also competitive. We also used MULTI-EURLEX to highlight the effect of temporal concept drift and the importance of chronological, instead of random, splits.

In future, we would like to examine alternative cross-lingual adaptation strategies (Pfeiffer et al., 2020, 2021) and distributionally robust optimization techniques (Sagawa et al., 2020; Koh et al., 2021) to address the temporal concept drift.

Ethics Statement

The dataset contains publicly available EU laws that do not include personal or sensitive information, with the exception of trivial information presented by consent, e.g., the names of the active presidents of the European Parliament, European Council, or other official administration bodies. The collected data is licensed under the Creative Commons Attribution 4.0 International licence (https://eur-lex.europa.eu/content/legal-notice/legal-notice.html). MULTI-EURLEX covers 23 languages from seven language families (Germanic, Romance, Slavic, Uralic, Baltic, Semitic, Hellenic). This does not imply that no other languages are spoken in EU countries, although EU laws are not translated to other languages (https://europa.eu/european-union/about-eu/eu-languages_en). We also provide a detailed Dataset Card (Gebru et al., 2018) for the MULTI-EURLEX dataset in Appendix D.

Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD)[14] under File No. 0175-00011A. This research has been also co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (T2EΔK-03849). We would like to thank Prodromos Malakasiotis, Nikolaos Aletras, Anders Søgaard, and Yoav Goldberg for providing valuable feedback, as well as Reviewer #3 for a particularly thorough review and feedback.

References

Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors. 2019. *Proceedings of the Natural Legal Language Processing Workshop 2019*. Minneapolis, Minnesota.

Nikolaos Aletras et al. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *Proceedings of the 31st International Conference on Legal Knowledge and Information Systems (JURIX)*, Groningen, The Netherlands.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jimmy Lei Ba, Jamie R. Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. In *NIPS 2016 Deep Learning Symposium*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020a. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020b. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Online.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium.

[14] https://innovationsfonden.dk/en

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume abs/1810.04805.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China.

Task force of the European Commision. 2012. Special Eurobarometer 386: Europeans and their Languages. EU Directorate-General for Communication.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. 2021. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. In *9th International Conference on Learning Representations (ICLR 2021)*, Online.

Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 915–926, Barcelona, Spain (Online).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Teresa Gonalves and Paulo Quaresma. 2010. Multilingual text classification through combination of monolingual classifiers. In *CEUR Workshop*, volume 605.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodríguez Penagos, and Marta Villegas. 2021. Spanish language models. *CoRR*, abs/2107.07253.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Mi-young Kim, Ying Xu, and Randy Goebel. 2015. A Convolutional Neural Network in Legal Question Answering. *Ninth International Workshop on Juris-informatics (JURISIN)*.

Diederik P. Kingma and Jim Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and Phil Blunsom. 2021. Pitfalls of static language modelling. *CoRR*, abs/2102.01951.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online.

Eneldo Loza Mencia and Johannes Fürnkranz. 2007. Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Proceedings of the LWA 2007*, pages 126–132.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1101–1111.

Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

Son Nguyen, Le-Minh Nguyen, Satoshi Tojo, Ken Satoh, and Akira Shimazu. 2018. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artificial Intelligence and Law*, 26:1–31.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Amir Rosenfeld and John K. Tsotsos. 2019. Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. In *16th Conference on Computer and Robot Vision (CRV 2021)*, Kingston, ON, Canada.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: towards more challenging and nuanced multilingual evaluation. *CoRR*, abs/2104.07412.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *8th International Conference on Learning Representations (ICLR 2020)*.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – czech bert-like model for language representation. *arXiv preprint arXiv:2103.13031*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about

random splits. In *Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *31th Annual Conference on Neural Information Processing Systems*, USA.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for finnish. *CoRR*, abs/1912.07076.

Bernhard Waltl, Johannes Muhr, Ingo Glaser, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2017. Classifying legal norms with active machine learning. In *30th International Conference on Legal Knowledge and Information Systems (JURIX)*, pages 11–20, Luxembourg.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *CoRR*, abs/2106.10199.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online.

A Implementation Details

A.1 Hyper-parameter Tuning

Similarly to previous work with pretrained TRANSFORMER-based models, we conduct grid-search to find the optimal learning rate per method considering classification performance on development data. We use early stopping on development data, if there is no improvement of mRP for five epochs. For end-to-end and partial (first N blocks frozen) fine-tuning, we search in {4e-5, 3e-5, 2e-5, 1e-5}, as suggested by Devlin et al. (2019); we also include in the search an even smaller learning rate (1e-6) as proposed by Conneau et al. (2020). For all (native) BERT models and XLM-ROBERTA, 3e-5 provided the best development results. For MT5 we search in {1e-3, 1e-4, 3e-5}, while Xue et al. (2021) proposed a fixed learning rate of 1e-3; in our case, 1e-4 provided the best development results. When we use adapter modules, BITFIT, or LNFIT, we search in {1e-3, 1e-4, 3e-5}, following Houlsby et al. (2019); again 1e-4 gave the best development results. While Houlsby et al. (2019) reported stable results across learning rates, in our case 1e-3 led to very unstable training with terrible performance.

For the bottleneck in adapter modules, where we have to select the number of hidden units (K), we search in {64, 128, 256, 384, 512}; 256 gave us the best development results, while the rest are comparable (Table 9).

K	Params	**en** (Src)	**All**
64	2.4M	72.5	57.8
128	4.8M	72.9	59.6
256	9.5M	72.5	60.2
384	14.2M	73.5	58.7
512	18.9M	72.6	56.6

Table 9: Development results for different values of K in adapter modules. We show mRP results (%) on English development data (Src), and development mRP averaged over all 23 languages (All). We also report the number of trainable parameters (in millions).

A.2 Other Technical Details

Given the large length of the documents (450 tokens on average), presented in Table 1, we truncate the documents, if needed, and use the first 512 tokens (sub-words) across all methods. Chalkidis et al. (2019) experimented with RNN-based methods using the full or truncated (up to 512 tokens) documents of EURLEX57K (English only, 7.4k labels), reporting almost identical results, i.e., the first 512 tokens of a document are adequate.

We also use label smoothing (Szegedy et al., 2016) ($\alpha = 0.2$) for levels 1–3, as we found it improves cross-lingual transfer in preliminary experiments. Label smoothing severely harms performance in experiments with the original label assignment (full label set with 7.4k labels).

All experiments ran on an NVIDIA DGX-1 station

with 8 NVIDIA V100 16GB GPU cards, although each experiment (model) was running on a single GPU card at a time. In Table 10, we report the average run-time per training experiment.

Model (Strategy)	L_{train}	Avg. run-time
NATIVE-BERT (Full)	1	10h
XLM-ROBERTA (Full)	1	12h
XLM-ROBERTA (Full)	23	12h
» First 3 blocks frozen	1	11h
» First 6 blocks frozen	1	8h
» First 9 blocks frozen	1	7h
» All 12 blocks frozen	1	3h
» Adapter modules	1	10h
» Adapter modules	23	15h
» BITFIT	1	18h
» LNFIT	1	11h

Table 10: Average training run-time across methods for Level 3 (575). L_{train} is the number of training languages. While BITFIT and LNFIT only tune a very small fraction (approx. 1-4× 1e-3%) of the parameters, training takes equal or longer, because the models are trained for more epochs and also there are trainable parameters as low as in the first TRANSFORMER block.

B Decoder Variants of MT5

generative: In preliminary experiments, we experimented with MT5's generative fine-tuning, using both the encoder and the decoder, as proposed by Xue et al. (2021). First, we ordered alphabetically the labels $(l_1, l_2, \ldots, l_N)$ by their identifiers. At each timestep, the decoder generates a token representing a label; i.e., we generate `[id_1]` for l_1, `[id_2]` for l_2, etc. Similarly to MT5 v1.1, we use a new (randomly initialized) classification layer (for a fixed vocabulary representing the N labels) to generate the output token at each timestep, based on the hidden state of the decoder. The entire model is trained to predict the labels in alphabetic order (in terms of `[id_i]`, where $i = 1, 2, \ldots, N$), but we ignore the order of the generated (predicted) labels during evaluation, to not penalize the model for not respecting the order. To rank the predicted labels when computing MRP, we use the probabilities (over the output vocabulary) assigned by the decoder to the corresponding generated tokens. We call *generative* this (original) version of MT5.

decode-cls: We also examined another MT5 variant, where again both the encoder and the decoder are used. In this variant, *decode-cls*, we feed the decoder with a single `[cls]` token (only one decoding timestep); by contrast, in *generative* the decoder performs multiple timesteps, and at each one it is fed with the output generated so far (or the corresponding gold output up to the previous timestep during training). In effect, in its single timestep, the decoder of *decode-cls* iteratively (at each decoder block) performs cross-attention over the encoder's output, using an updated query (`[cls]` representation). We pass the final representation of the decoder's `[cls]` to the same classification layer we use in the encoder-only variant of MT5 (Section 4). Both *decode-cls* and *generative* use 12 encoder and 12 decoder blocks, 391M parameters.

MT5 variant	Params	Train Time	**en** (Src)	**All**
first-pool	277M	32e / 25h	72.4	**55.6**
last-pool		18e / 14h	**72.8**	48.8
generative	391M	3e / 3h	2.5	2.5
decode-cls		21e / 19h	72.7	52.8
XLM-ROBERTA	278M	22e / 12h	73.1	50.4

Table 11: Comparing MT5 variants. The first two variants use only the encoder; the latter two use both the encoder and the decoder. We show MRP results (%) on English development data (Src), and development MRP averaged over all 23 languages (All). We also report the number of trainable parameters (in millions), and training time in epochs (e) and hours (h).

first-pool, last-pool: Finally, we examine another encoder-only variant of MT5, *last-pool*, in addition to the encoder-only variant of Section 4, which we now call *first-pool* to highlight the difference between them. In *last-pool*, we use the encoder's top-level representation of the `</s>` special token of MT5, which is always at the end of the input, to represent the document. Since the position of `</s>` is not always the same, its representation is also affected by its positional embedding. By contrast, in *first-pool* the `[cls]` token is always first, hence its positional embedding does not vary.

Table 11 reports results on development data. As expected, the *generative* version of MT5 performs terribly (MRP 2.5), as the model tries to learn an unnecessary label ordering; in fact the model cannot learn and stops training after five epochs due to early stopping. By contrast, the *decode-cls* variant, which feeds the decoder only with the `[cls]` token and uses its output embedding, has comparable performance with the encoder-only variants (*first-pool*, *last-pool*). It uses, however, approximately 40% more parameters, because of the additional cross-attention layers in the decoder blocks.

Both encoder-only variants of MT5 are comparable with XLM-ROBERTA (English MRP approx. 73; the All MRP scores are also comparable or bet-

Language		Model	Publication	Pretraining Corpora
English	**(en)**	`bert-base-uncased`	(Devlin et al., 2019)	Wikipedia + Books
Danish	**(da)**	`DJSammy/bert-base-danish-uncased_BotXO,ai`	-	Wikipedia + Web + Subtitles
German	**(de)**	`deepset/gbert-base`	(Chan et al., 2020)	Wikipedia + OSCAR + OPUS
Dutch	**(nl)**	`pdelobelle/robbert-v2-dutch-base`	(Delobelle et al., 2020)	Wikipedia + Books + News
Swedish	**(sv)**	`KB/bert-base-swedish-cased`	-	Wikipedia + Books + News
Spanish	**(es)**	`BSC-TeMU/roberta-base-bne`	(Gutiérrez-Fandiño et al., 2021)	Web
French	**(fr)**	`camembert-base`	(Martin et al., 2020)	OSCAR
Italian	**(it)**	`dbmdz/bert-base-italian-uncased`	-	Wikipedia + OPUS
Portuguese	**(pt)**	`neuralmind/bert-base-portuguese-cased`	(Souza et al., 2020)	Web
Czech	**(cs)**	`UWB-AIR/Czert-B-base-cased`	(Sido et al., 2021)	Wikipedia + Web + News
Romanian	**(ro)**	`dumitrescustefan/bert-base-romanian-uncased-v1`	-	Wikipedia + OSCAR + OPUS
Polish	**(pl)**	`dkleczek/bert-base-polish-uncased-v1`	-	Wikipedia
Estonian	**(et)**	`tartuNLP/EstBERT`	-	Web
Finish	**(fi)**	`TurkuNLP/bert-base-finnish-uncased-v1`	(Virtanen et al., 2019)	Web + News
Hungarian	**(hu)**	`SZTAKI-HLT/hubert-base-cc`	(Nemeskey, 2020)	Wikipedia + OSCAR
Greek	**(el)**	`nlpaueb/bert-base-greek-uncased-v1`	(Koutsikakis et al., 2020)	Wikipedia + OSCAR

Table 12: Monolingual (native) BERT models used. We also report the training corpora used to pretrain each model.

ter). These results show that the encoder of MT5 can be used alone (without the decoder) for text classification, similarly to TRANSFORMER-based encoder-only models (Devlin et al., 2019; Liu et al., 2019), despite its text-to-text generative pretraining, unlike the generative fine-tuning proposed by the creators of MT5 (Xue et al., 2021).

C Monolingual BERT Models

Table 12 lists all native BERT models used in the experiments of Section 6.1 in the one-to-one set-up. All models are hosted by Hugging Face (`https://huggingface.co/models`). All models follow the BASE configuration with 12 layers of stacked TRANSFORMERs, each with $D_h = 768$ hidden units and 12 attention heads. We use case sensitive models, when available. We cannot guarantee the quality of the different models, as they come from different sources (organizations or individuals), although we tried to select the best possible options, i.e., those trained on more data for a longer period, in case there were many alternatives. We found 16 monolingual models; we found no monolingual models for Bulgarian, Slovak, Croatian, Slovene, Lithuanian, Latvian, Maltese.

Most monolingual BERT models use a vocabulary of approx. 30k sub-words and have approx. 110M parameters in total (24M for embeddings and 86M for TRANSFORMER blocks), while XLM-ROBERTA has a much larger vocabulary of 250k sub-words to support 100 languages and 278M parameters (192M for embeddings and 86M for TRANSFORMER blocks). Similarly, MT5 uses a vocabulary of equal size, thus its encoder has 86M parameters, while its decoder has 120M parameters; as in the work of Vaswani et al. (2017), the decoder TRANSFORMER blocks of MT5 have more parameters than the encoder blocs, as they use additional cross-attention layers. Based on the aforementioned details, the encoder's capacity is almost identical across the examined models.

D Dataset Card for MULTI-EURLEX

D.1 Dataset Description

Documents: MULTI-EURLEX comprises 65k EU laws (published 1958–2016) in 23 official EU languages (Table 1). Each EU law has been annotated with EUROVOC concepts (labels) by EU's Publications Office. Each EUROVOC label ID is associated with a *label descriptor*, e.g., ⟨60, 'agri-foodstuffs'⟩, ⟨6006, 'plant product'⟩, ⟨1115, 'fruit'⟩. The descriptors are also available in the 23 languages.

Languages: The EU has 24 official languages. When new members join the EU, the set of official languages usually expands, unless the new languages are already included. MULTI-EURLEX covers 23 languages from seven language families (Germanic, Romance, Slavic, Uralic, Baltic, Semitic, Hellenic). EU laws are published in all official languages, except Irish, for resource-related reasons.[15] This wide coverage makes MULTI-EURLEX a valuable testbed for cross-lingual transfer. All languages use the Latin script, except for Bulgarian (Cyrillic script) and Greek. Several other languages are also spoken in EU countries. The EU is home to over 60 additional indigenous regional or minority languages, e.g., Basque, Catalan, Frisian, Saami, and Yiddish, among others, spoken by approx. 40 million people, but these additional languages are not considered official (in terms of EU), and EU laws are not translated to them.

Annotation: All the documents of the dataset have been annotated by the Publications Office

[15] `https://europa.eu/european-union/about-eu/eu-languages_en`

of EU (https://publications.europa.eu/en) with multiple concepts from EUROVOC (http://eurovoc.europa.eu/). EUROVOC has eight levels of concepts. Each document is assigned one or more concepts (labels). If a document is assigned a concept, the ancestors and descendants of that concept are typically not assigned to the same document. The documents were originally annotated with concepts from levels 3 to 8. We augmented the annotation with three alternative sets of labels per document, replacing each assigned concept by its ancestor from level 1, 2, or 3, respectively. Thus, we provide four sets of gold labels per document, one for each of the first three levels of the hierarchy, plus the original sparse label assignment.[16]

Data Split and Concept Drift: MULTI-EURLEX is *chronologically* split in training (55k), development (5k), test (5k) subsets, using the English documents. The test subset contains the same 5k documents in all 23 languages (Table 1).[17] For the official languages of the seven oldest member countries, the same 55k training documents are available; for the other languages, only a subset of the 55k training documents is available (Table 1). Compared to EURLEX57K (Chalkidis et al., 2019), MULTI-EURLEX is not only larger (8k more documents) and multilingual; it is also more challenging, as the chronological split leads to temporal real-world *concept drift* across the training, development, test subsets, i.e., differences in label distribution and phrasing, representing a realistic *temporal generalization* problem (Lazaridou et al., 2021). Søgaard et al. (2021) showed this setup is more realistic, as it does not over-estimate real performance, contrary to random splits (Gorman and Bedrick, 2019).

Supported Tasks: MULTI-EURLEX can be used for legal topic classification, a multi-label classification task where legal documents need to be assigned concepts reflecting their topics. MULTI-EURLEX supports labels from three different granularities (EUROVOC levels). More importantly, apart from monolingual (*one-to-one*) experiments, it can be used to study cross-lingual transfer scenarios, including *one-to-many* (systems trained in one language and used in other languages with no training data), and *many-to-one* or *many-to-many* (systems jointly trained in multiple languages and used in one or more other languages).

Data Fields: The following data fields are provided for all documents of MULTI-EURLEX:

- 'celex_id': (**str**) The official ID of the document. The CELEX number is the unique identifier for all publications in both EUR-LEX and CELLAR, the EU Publications Office's common repository of metadata and content.
- 'publication_date': (**str**) The publication date of the document.
- 'text': (**dict[str]**) A dictionary with (key, value) pairs, where the key is the 2-letter ISO code of each language and the value is the content of each document in this language.
- 'eurovoc_concepts': (**dict[List[str]]**) A dictionary with (key, value) pairs, where the key is the label set (level 1–3) and the value is a list of the relevant EUROVOC concepts (labels).

D.2 Initial Data Collection and Normalization

The original data are available at the EUR-LEX portal (https://eur-lex.europa.eu) in unprocessed formats (HTML, XML, RDF). The documents were downloaded from the EURLEX portal in HTML. The relevant EUROVOC concepts were downloaded from the SPARQL endpoint of the Publications Office of EU (http://publications.europa.eu/webapi/rdf/sparql). We stripped HTML mark-up to provide the documents in plain text format. We inferred the labels for EUROVOC levels 1–3, by backtracking the EUROVOC hierarchy branches, from the originally assigned labels to their ancestors in levels 1–3, respectively.

D.3 Personal and Sensitive Information

The dataset contains publicly available EU laws that do not include personal or sensitive information, with the exception of trivial information presented by consent, e.g., the names of the current presidents of the European Parliament and European Council, and other administration bodies.

D.4 Licensing Information

We provide MULTI-EURLEX with the same licensing as the original EU data (CC-BY-4.0):

The Commission's document reuse policy is based on Decision 2011/833/EU. Unless otherwise specified, you can re-use the legal documents published

[16] Levels 4 to 8 cannot be used independently, as many documents have gold concepts from the third level; thus many documents will be mislabeled, if we discard level 3.

[17] The development subset also contains the same 5k documents in 23 languages, except Croatian. Croatia is the most recent EU member (2013); older laws are gradually translated.

in EUR-LEX *for commercial or non-commercial purposes.*

The copyright for the editorial content of this website, the summaries of EU *legislation and the consolidated texts, which is owned by the* EU*, is licensed under the Creative Commons Attribution 4.0 International licence. This means that you can re-use the content provided you acknowledge the source and indicate any changes you have made.*

Source: https://eur-lex.europa.eu/content/legal-notice/legal-notice.html

See also: https://eur-lex.europa.eu/content/help/faq/reuse-contents-eurlex.html

E More Detailed Results

For completeness, in Table 14 we present detailed results across all 23 languages for XLM-ROBERTA fine-tuned end-to-end or using the alternative adaptation strategies in the *one-to-many* setting for English. We observe that (a) native BERT models have the best results in 12 out of 15 languages; (b) XLM-ROBERTA trained in a monolingual (one-to-one) setting has competitive results; and (c) fine-tuning with adapter modules leads to the best overall results in cross-lingual transfer and in the many-to-many setting.

Tables 15–16 show the results when fine-tuning end-to-end or using adapters, considering each one of the 23 languages as a source language in a *one-to-many* setting.

Table 18 shows XLM-ROBERTA results for all EUROVOC levels across all 23 languages.

Table 17 reports detailed results across all 23 languages for the alternative adaptation strategies using the *first-pool* MT5 variant.

Similarly to Table 4, Table 13 shows the effects of temporal concept drift in the performance of XLM-ROBERTA, for Level 3 with 567 labels.

Data Split	Training	Development	Test
Random	**93.0**	**80.9**	**80.3**
Chronological	92.8	73.1	67.4

Table 13: Results of MULTI-EURLEX for level 3 (567 labels) with XLM-ROBERTA using a *random* or *chronological* split. Here the model is fine-tuned and tested on English data only (*one-to-one*).

		GERMANIC				ROMANCE					SLAVIC						URALIC			BALTIC				
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																								
NATIVE-BERT	67.7	65.5	68.4	66.7	68.5	68.5	67.6	67.4	67.9	67.4	-	66.7	-	67.2	-	-	67.7	67.8	66.0	-	-	67.8	-	67.4
XLM-ROBERTA	67.4	66.7	67.5	67.3	66.5	66.4	67.8	67.2	67.4	67.0	66.1	66.7	61.7	65.0	64.8	66.7	65.5	66.5	65.7	66.2	66.7	65.8	62.9	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-	0.0	-	-2.2	-	-	-2.2	-1.3	-0.3	-	-	-2.0	-	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																								
End-to-end fine-tuning	67.4	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	51.2	49.6	49.6	46.9	49.3	49.9	48.8	46.4	45.2	49.7	46.4	33.3	20.4	49.3
First three blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	56.1	54.3	52.8	53.7	53.0	51.4	51.0	52.1	49.7	51.3	50.1	42.4	20.3	53.0
First six blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	55.8	52.6	49.2	51.9	50.8	49.3	47.3	48.7	45.0	48.5	49.9	39.6	22.0	51.7
First nine blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	56.7	54.2	50.7	53.4	54.4	48.7	48.8	50.4	46.2	51.6	50.5	44.5	21.4	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	23.1	24.3	19.9	22.8	26.0	19.8	22.8	21.9	20.2	22.9	21.4	19.0	14.2	22.2
Adapters layers	67.3	61.5	59.3	57.8	59.5	60.3	61.0	60.4	58.8	58.5	59.2	56.8	56.9	57.5	57.0	53.5	55.3	55.6	53.1	55.2	52.4	46.1	27.4	56.1
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	55.6	54.8	55.1	54.0	52.8	57.9	51.2	54.8	52.3	52.5	51.8	42.1	22.7	53.7
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																								
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.7	65.7	65.8	65.6	65.7	65.8	65.2	65.8	65.6	65.7	65.8	65.1	62.3	65.7
Adapters layers	**67.3**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.6**	**67.0**	**67.0**	**66.2**	**66.2**	**66.8**	**65.5**	**66.6**	**65.7**	**65.8**	**66.7**	**65.7**	**61.6**	**66.4**

Table 14: Test results for XLM-ROBERTA in cross-lingual classification at level 3 (567 labels). We show mRP (%) for each one of the 23 languages, and mRP averaged over all 23 languages.

	GERMANIC					ROMANCE					SLAVIC						URALIC			BALTIC				
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
en	67.4	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	51.2	49.6	49.6	46.9	49.3	49.9	48.8	46.4	45.2	49.7	46.4	33.3	20.4	49.3
da	55.6	66.7	54.5	53.6	58.3	50.8	50.2	48.6	47.7	47.1	49.7	46.9	47.1	44.1	46.6	47.3	46.5	42.4	43.1	47.0	43.1	33.6	17.7	47.3
de	55.3	56.9	67.5	52.9	55.9	50.3	49.4	48.6	46.2	45.3	48.1	49.0	47.9	46.5	50.2	46.5	49.6	44.5	44.0	43.5	40.5	34.9	17.5	47.4
nl	55.9	56.5	55.2	67.3	53.1	50.1	50.8	50.0	48.1	47.5	49.3	49.8	49.5	48.7	50.6	46.8	45.6	45.3	42.4	47.1	44.2	35.9	18.2	48.2
sv	58.3	61.5	55.0	52.5	66.5	52.2	51.1	50.7	49.2	49.9	51.7	50.0	50.2	48.4	50.4	49.3	47.5	47.9	48.3	49.8	45.0	37.0	18.8	49.6
ro	57.7	53.2	51.9	50.5	52.5	66.4	59.8	57.4	56.2	54.1	53.4	50.7	50.9	51.2	50.6	52.6	51.0	46.2	49.2	50.2	45.9	40.4	18.3	50.9
es	58.6	53.8	49.8	48.3	52.6	59.3	67.8	58.9	59.3	61.5	52.5	47.8	49.3	48.6	47.4	46.8	47.2	41.6	41.4	44.2	43.4	36.4	19.3	49.4
fr	60.6	55.0	52.1	54.1	53.9	58.4	60.3	67.2	57.8	58.5	53.9	50.9	52.8	51.0	50.8	50.5	49.6	46.0	45.4	47.8	44.6	38.1	22.1	51.4
it	56.7	52.3	46.1	45.4	50.2	56.5	60.7	56.9	67.4	56.6	51.3	46.2	47.9	43.7	47.5	47.6	44.1	40.6	42.0	42.8	40.1	33.9	22.8	47.8
pt	59.0	54.5	50.5	48.6	52.2	57.9	63.0	59.2	58.8	67.0	51.9	47.5	50.2	48.5	49.0	51.1	49.5	45.2	42.3	46.3	43.2	35.9	20.7	50.1
bg	55.8	51.5	51.1	48.3	50.5	53.5	52.0	51.2	47.9	48.7	66.1	50.4	53.3	50.6	50.2	51.0	48.3	46.4	40.7	50.1	45.6	39.6	14.7	48.6
cs	53.1	52.4	50.5	48.0	51.2	49.1	52.0	48.9	48.9	49.1	51.7	66.7	55.9	51.7	61.7	55.8	50.5	46.3	44.2	49.4	44.7	36.0	16.1	49.3
hr	50.4	50.6	49.1	48.1	50.8	52.8	50.5	49.9	50.6	47.9	51.4	51.8	61.7	51.4	53.3	54.6	47.0	49.1	46.8	47.9	46.5	37.0	19.6	48.6
pl	53.6	50.8	48.5	47.7	50.2	51.3	52.1	49.1	45.4	49.1	52.2	51.7	51.6	65.0	51.9	49.8	45.6	42.1	39.5	44.4	41.3	35.3	13.1	47.0
sk	53.8	53.2	52.4	49.7	51.2	50.9	52.1	50.8	49.1	48.4	53.6	60.4	54.4	53.5	64.8	54.1	49.4	47.2	44.9	49.2	45.0	38.1	16.0	49.7
sl	54.2	53.4	49.5	50.0	52.4	52.2	53.1	51.5	51.8	49.9	53.3	54.6	58.1	51.0	53.3	66.7	48.7	47.7	45.5	47.8	45.8	38.7	15.0	49.7
hu	51.9	49.7	47.5	42.9	46.9	48.9	48.6	47.1	45.2	43.8	48.8	47.0	46.8	46.5	47.0	46.3	65.5	44.3	43.9	44.2	41.3	33.4	14.2	45.3
fi	47.6	48.5	43.7	41.9	48.2	44.7	43.3	42.0	43.8	38.7	42.7	42.6	45.2	38.5	43.3	45.6	43.9	66.5	45.2	43.7	38.7	30.3	11.2	42.6
et	51.2	49.5	44.8	44.6	48.7	45.8	48.2	43.9	44.6	42.6	43.4	44.6	44.9	42.6	44.6	45.4	46.1	46.2	65.7	45.2	43.4	32.2	14.7	44.5
lt	54.3	48.4	46.0	41.9	49.3	49.8	48.9	47.4	46.5	43.4	48.2	46.0	48.4	46.7	47.2	49.3	45.2	41.2	43.9	66.2	49.0	32.2	16.9	45.9
lv	51.8	49.4	45.9	48.1	49.3	51.3	50.6	49.8	47.8	48.4	51.1	48.9	48.9	49.2	49.6	49.5	49.0	44.3	48.3	53.0	66.7	38.1	18.8	48.1
el	41.2	37.8	34.3	33.3	38.0	38.2	42.3	38.5	39.6	41.3	39.3	34.5	38.8	32.1	36.8	35.6	36.8	31.3	28.9	34.7	30.3	65.8	9.6	36.5
mt	27.5	28.4	23.2	25.3	25.5	25.6	25.3	24.8	22.9	23.9	24.1	22.2	22.0	21.9	23.0	22.1	23.6	21.0	22.9	23.9	23.2	19.6	62.9	25.4

Table 15: Test results (mRP, %) for Level 3 (567 labels) with XLM-ROBERTA, when fine-tuning end-to-end in one language (source, rows) and testing in all languages (columns).

	GERMANIC					ROMANCE					SLAVIC						URALIC			BALTIC				
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
en	66.8	61.5	59.3	57.8	59.5	60.3	61.0	60.4	58.8	58.5	59.2	56.8	56.9	57.5	57.0	53.5	55.3	55.6	53.1	55.2	52.4	46.1	27.4	56.1
da	62.0	66.6	60.1	59.6	62.4	58.4	58.7	57.1	54.5	56.4	58.1	57.2	56.9	55.4	57.9	55.0	56.6	55.9	52.5	55.8	52.7	46.5	21.2	55.5
de	60.7	61.0	67.1	58.5	59.8	57.7	57.8	56.9	56.1	55.5	56.7	56.1	55.8	55.9	56.2	54.4	55.8	53.5	51.7	51.3	51.3	44.8	24.5	54.8
nl	61.6	62.0	59.9	67.4	59.2	58.7	58.5	58.1	55.0	56.1	57.6	58.2	57.4	56.1	58.2	54.0	55.7	53.7	50.8	54.3	52.0	45.4	23.4	55.4
sv	61.8	63.8	60.4	58.2	66.9	57.1	57.8	56.5	55.4	55.8	57.3	56.8	55.3	55.1	57.1	55.3	53.7	55.0	52.5	54.4	51.5	44.9	22.7	55.0
ro	62.0	59.8	59.3	58.1	58.0	67.0	60.9	59.8	58.9	58.8	59.5	56.5	56.8	57.4	56.8	58.2	56.1	55.0	55.9	56.8	55.3	48.4	29.2	56.7
es	62.8	59.2	56.5	57.3	57.1	62.7	67.7	61.7	61.8	63.0	58.3	57.0	55.9	56.6	57.5	53.1	54.1	51.8	51.1	50.6	52.3	45.3	23.1	55.5
fr	63.7	60.8	60.6	58.6	58.0	61.6	63.0	67.4	60.0	60.3	59.7	57.9	59.2	58.2	58.9	55.4	56.9	55.2	51.7	53.6	53.1	47.3	31.4	57.1
it	61.1	58.2	53.4	57.0	55.2	60.1	61.5	60.8	67.4	60.6	54.9	52.9	54.2	53.6	51.5	54.2	51.4	53.6	51.0	52.4	51.9	47.7	26.5	54.4
pt	63.9	59.1	57.1	57.5	56.8	61.4	64.9	60.6	61.2	67.6	59.2	56.0	57.9	56.1	56.6	55.4	54.6	52.6	49.6	53.4	53.0	47.1	24.4	55.9
bg	61.2	58.3	58.5	56.3	57.7	59.0	57.4	57.5	54.7	56.1	66.9	58.3	58.7	58.4	58.5	53.7	56.6	53.3	51.2	53.8	50.9	47.8	23.3	55.1
cs	59.8	60.1	58.0	56.8	56.9	57.5	57.8	58.5	56.6	56.1	59.0	67.2	60.5	59.4	64.6	59.0	57.2	54.8	54.8	56.1	55.4	46.9	25.7	56.5
hr	56.0	52.4	52.5	51.7	54.8	54.3	53.8	54.5	53.0	51.6	55.4	54.7	62.4	52.7	55.6	55.3	49.9	50.8	49.4	52.3	50.1	41.6	24.6	51.7
pl	59.2	58.8	57.4	55.8	56.7	59.2	57.3	57.4	53.7	56.4	58.8	58.5	57.6	66.5	59.7	56.0	55.1	54.9	48.6	54.3	53.2	45.6	20.6	54.8
sk	60.0	59.5	58.0	57.7	58.3	59.3	58.6	57.8	55.5	54.5	59.7	62.5	60.4	58.6	66.6	59.6	56.3	55.3	55.7	55.4	55.9	46.4	23.4	56.3
sl	60.8	59.3	57.9	57.6	59.2	59.1	59.1	58.0	56.1	56.7	60.5	59.5	62.5	58.2	60.0	66.0	56.1	56.5	54.2	56.0	55.6	48.3	24.8	56.6
hu	61.0	59.6	59.7	57.6	57.5	57.6	59.3	56.9	55.1	55.2	59.2	57.9	57.5	56.9	56.9	56.0	66.2	55.9	54.4	54.9	54.3	47.9	23.7	55.7
fi	58.6	57.3	54.7	53.8	56.5	56.8	56.8	53.9	54.5	54.2	54.4	54.9	53.3	53.0	53.2	53.2	55.2	66.8	54.4	51.0	48.8	43.1	21.9	53.1
et	58.8	57.0	54.8	54.4	57.4	57.0	56.0	54.1	54.4	52.8	55.8	55.0	55.6	54.4	55.3	53.3	57.1	54.9	66.2	54.6	53.7	44.0	21.0	53.8
lt	59.2	57.6	54.1	53.7	56.6	56.2	55.1	56.7	54.1	51.4	56.8	55.4	53.3	54.4	56.0	53.4	55.0	55.2	52.2	67.0	56.5	43.8	27.0	53.9
lv	58.6	57.1	56.4	53.7	57.7	56.8	56.4	54.7	54.3	53.5	57.2	55.4	55.0	55.3	55.9	54.9	55.5	55.2	55.2	58.1	66.5	45.4	22.3	54.4
el	46.8	48.9	47.8	49.5	49.3	48.8	52.0	50.6	48.9	49.4	51.0	50.1	48.6	48.4	49.9	47.7	48.2	46.2	41.5	44.2	45.1	65.5	20.6	47.8
mt	43.7	44.3	40.7	39.7	40.7	42.8	44.5	42.6	45.3	42.0	38.7	39.1	40.2	38.4	40.0	39.1	38.6	35.2	35.9	40.5	39.7	32.8	63.9	41.2

Table 16: Test results (mRP, %) for Level 3 (567 labels) with XLM-ROBERTA fine-tuned with adapter modules in one language (source, rows) and testing in all languages (columns).

		GERMANIC				ROMANCE					SLAVIC						URALIC			BALTIC				
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
One-to-many (Fine-tune MT5 *only* in English, test in all languages, with alternative adaptation strategies.)																								
End-to-end fine-tuning	67.4	59.5	58.0	57.1	58.9	58.5	58.3	59.6	54.9	54.8	56.2	52.3	40.7	52.2	50.9	53.2	52.6	51.1	51.3	53.3	51.4	43.5	39.5	53.7
First 3 blocks frozen	67.4	**62.6**	60.0	**60.5**	**61.0**	61.4	61.3	60.9	**60.0**	58.2	**58.8**	56.1	45.4	54.9	56.4	56.9	**54.5**	54.2	55.4	56.2	56.4	46.7	44.0	56.9
First 6 blocks frozen	66.3	61.8	59.3	61.1	61.7	61.0	61.5	61.7	58.9	59.5	60.9	57.9	48.5	57.8	57.9	59.4	56.2	58.7	59.2	58.9	57.1	50.9	46.3	58.4
First 9 blocks frozen	**68.0**	61.9	**60.8**	59.1	**61.0**	**63.1**	**63.2**	**63.7**	59.1	**61.8**	58.7	**58.4**	**47.1**	**56.8**	**58.2**	**59.4**	54.0	**55.6**	**57.2**	**58.1**	**57.5**	**48.5**	**48.7**	**58.3**
Adapters layers	66.3	53.0	48.8	47.1	49.3	48.3	53.8	52.9	49.8	48.4	45.5	41.7	28.2	43.1	35.9	38.0	41.2	42.5	35.8	38.4	40.4	37.2	27.5	44.0

Table 17: Test results for MT5 (first-pool) in cross-lingual classification at level 3 (567 labels). We show mRP (%) for each one of the 23 languages, and mRP averaged over all 23 languages.

EUROVOC			GERMANIC				ROMANCE					SLAVIC						URALIC			BALTIC				
Label Set	(#Labels)	**en**	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
One-to-many (Fine-tune XLM-ROBERTA end-to-end *only* in English, test in all languages.)																									
Level 1	(21)	83.2	78.0	78.5	75.9	75.8	77.4	78.7	78.6	77.2	77.5	78.3	76.6	76.7	76.8	76.0	76.2	75.5	74.5	75.0	76.2	75.3	71.4	52.9	75.7
Level 2	(127)	73.6	64.1	61.2	58.8	61.0	64.5	65.6	63.0	61.8	60.3	61.1	59.0	59.1	58.8	58.1	60.5	58.5	55.4	55.4	57.8	55.1	47.2	31.3	58.7
Level 3	(567)	67.4	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	51.2	49.6	49.6	46.9	49.3	49.9	48.8	46.4	45.2	49.7	46.4	33.3	20.4	49.3
All	(7,390)	43.0	28.5	26.9	25.4	30.6	31.5	29.2	30.5	30.2	30.1	28.4	21.6	25.1	24.5	22.2	24.5	20.5	20.9	18.8	19.2	17.3	14.9	4.7	24.7
One-to-many (Fine-tune XLM-ROBERTA with adapter modules *only* in English, test in all languages.)																									
Level 1	(21)	83.1	80.3	79.1	78.9	77.9	80.3	78.0	78.6	79.1	78.4	79.7	77.7	77.5	77.5	77.6	77.9	76.0	76.2	76.5	76.7	78.5	75.3	54.8	77.2
Level 2	(127)	73.2	65.9	66.6	62.3	62.0	66.9	65.2	67.0	62.0	64.1	64.3	61.4	65.0	60.1	62.4	64.5	61.6	59.4	56.8	58.0	61.2	49.4	31.2	61.3
Level 3	(567)	66.8	61.5	59.3	57.8	59.5	60.3	61.0	60.4	58.8	58.5	59.2	56.8	56.9	57.5	57.0	53.5	55.3	55.6	53.1	55.2	52.4	46.1	27.4	56.1
All	(7,390)	42.8	37.0	32.7	34.5	36.2	36.7	35.7	33.5	33.4	36.7	35.7	35.0	33.0	32.7	34.5	32.6	31.6	30.1	31.6	32.2	30.2	27.3	13.4	33.0

Table 18: Test results of XLM-ROBERTA fine-tuned *end-to-end* or with adapters across all EUROVOC levels (label sets).

	GERMANIC					ROMANCE					SLAVIC						URALIC			BALTIC				
Family (Src)	**en**	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**bg**	**cs**	**hr**	**pl**	**sk**	**sl**	**hu**	**fi**	**et**	**lt**	**lv**	**el**	**mt**	All
Many-to-many (Fine-tune XLM-ROBERTA end-to-end in *all languages of the same family*, test in all languages.)																								
GERMANIC	67.9	67.8	67.6	66.1	67.4	61.1	60.2	60.3	58.2	58.9	60.6	58.0	61.2	58.1	58.4	59.4	57.0	56.3	52.9	57.9	56.5	45.9	26.4	58.4
ROMANCE	65.2	58.5	54.8	56.7	57.7	67.1	67.3	67.4	67.6	66.0	60.9	56.0	57.3	55.3	55.1	56.5	54.0	48.3	50.3	51.0	49.5	40.0	30.9	56.2
SLAVIC	63.8	60.7	62.0	61.6	60.5	62.9	61.5	61.6	59.7	59.3	67.4	66.6	67.1	65.6	65.0	65.6	57.7	59.2	58.6	60.7	59.0	49.0	23.2	59.9
URALIC	60.4	56.1	55.2	50.4	55.7	56.3	55.0	53.4	50.4	50.9	52.1	52.9	56.1	49.0	53.1	53.1	65.5	66.0	49.4	53.0	50.2	39.8	16.1	52.2
BALTIC	59.8	53.3	47.6	51.9	53.5	53.4	53.2	52.2	51.2	51.0	55.2	53.5	54.7	52.7	54.8	57.4	51.0	48.2	53.0	67.3	66.7	36.1	18.9	52.0
Many-to-many (Fine-tune XLM-ROBERTA with adapter modules in *all languages of the same family*, test in all languages.)																								
GERMANIC	67.6	67.5	67.4	66.8	67.0	65.3	65.2	62.7	60.3	62.1	64.2	64.0	62.3	63.5	64.0	62.1	62.1	62.4	59.7	59.6	62.2	52.4	30.4	61.8
ROMANCE	66.6	65.6	62.9	63.3	61.3	66.1	67.9	67.8	67.7	67.5	64.8	61.8	60.6	60.8	61.9	61.3	62.0	59.5	59.6	59.1	61.0	54.8	33.5	61.6
SLAVIC	63.6	63.0	62.7	61.5	64.0	62.9	61.8	62.4	57.9	59.1	66.0	66.4	64.8	66.2	66.3	66.5	61.6	61.4	56.0	61.6	60.6	53.1	33.6	61.0
URALIC	61.5	60.3	59.3	59.4	60.2	57.9	59.3	58.1	56.1	57.7	60.3	58.5	59.3	57.9	56.7	58.1	66.5	66.7	58.2	57.9	56.4	49.4	28.6	57.6
BALTIC	62.4	60.0	57.2	59.2	59.8	60.1	60.9	58.1	58.2	58.5	60.8	56.4	62.9	60.0	56.7	58.9	57.5	58.9	56.3	66.0	65.9	51.2	30.7	58.1

Table 19: Test results (mRP, %) for Level 3 (567 labels) with XLM-ROBERTA, when fine-tuning *end-to-end* or with adapters in *all languages of the same family (Src)*, test in all languages. We show mRP (%) for each one of the 23 languages, and mRP averaged over all 23 languages.