Legal document classification in one shot cross lingual transfer setting

Noémie Guibé
Student at ENSAE
noemie.guibe@ensae.fr

Abstract

A reproduction study using the MultiEURLEX dataset (1958–2016) as part of the Natural Language Processing course, this report investigates multilingual and zero-shot legal text classification based on the framework proposed by Chalkidis et al. We explore challenges in cross-lingual transfer—particularly for low-resource languages—by evaluating layer freezing, adapter-based fine-tuning, and a language-agnostic approach grounded in lexical cues. Beyond standard performance metrics, we assess training time and model size to better capture the trade-offs involved in developing efficient and inclusive legal NLP systems.

1 A challenging NLP problem

In machine learning and NLP, one recurring challenge is the need for a high-quality labeled training dataset to effectively train a model for classification or prediction tasks. The quality of this dataset plays a significant role in the performance of the model. Since the model's output is highly dependent on the data it is trained on, it must be regularly retrained with new data to stay relevant, especially for evolving topics.

However, training a model with large datasets can be costly. To address this, many models are now provided in a pre-trained state. This allows users, especially those without the resources to train a model from scratch, to still utilize the technology. Pre-trained models can later be fine-tuned or retrained on domain-specific data, depending on the application.

Despite these advancements, creating labeled datasets is time-consuming and relies heavily on human annotators. Increasing the size of training datasets isn't a straightforward solution, as it raises ethical concerns and is not financially feasible for all organizations or countries. This disparity is particularly evident between resource-rich languages, like English, and languages with limited or no available training data in the field of NLP. Therefore, the first key question is: how can we optimally utilize a labeled training set?

This is where zero-shot classification comes into play. Zero-shot classification allows a model, trained on one dataset, to solve problems in a test dataset with different characteristics. In this approach, the model can classify data into categories it has never encountered during training. By providing the model with a natural language prompt and task description, it can understand and perform the classification without the need for prior examples.

This concept is also central to transfer learning, which involves leveraging data and models available for high-resource languages (e.g., English) to solve tasks in low-resource languages. Despite the recent advancements in NLP, particularly with the development of large language models that achieve exceptional performance across a range of tasks, these benefits have mostly been realized by high-resource languages. The majority of languages continue to face significant challenges due to the scarcity of training data and computational resources.

Chalkidis et al. [2021] addressed this issue in their September 2021 paper titled "*MultiAU-RLEX:*

A Multilingual and Multi-label Legal Document Classification Dataset for Zero-shot Cross-lingual Transfer" . They propose using cross-lingual transfer, where documents labeled in one language are used to classify documents in other languages, within the context of legal NLP. Additionally, they provide access to a dataset designed to facilitate this approach, with the goal of enabling classifiers trained on resource-rich languages to be applied to languages with fewer or no available training instances.

2 State of the art

Legal NLP is an emerging field focused on tasks such as legal judgment prediction, legal topic classification, legal question answering, and contract understanding, among others. One area that had not been explored in legal NLP until recently is cross-lingual transfer, which involves using resources from one language to solve tasks in another. The authors of the study decided to combine these two domains—cross-lingual transfer and legal NLP—to advance the field.

The data used in this study consists of 65,000 European Union laws, translated into 23 official EU languages. The primary goal is to leverage legal documents labeled in one language to train models that can classify documents in other languages. This is particularly challenging in zero- or one-shot settings due to the complexity of legal language and the presence of fine-grained, hierarchical, and overlapping label structures.

State-of-the-art multilingual transformer models have been the foundation for tackling these challenges, such as:

- XLM-RoBERTa (Conneau et al. [2020]): it is A BERT-style encoder-only model trained on 2.5TB of filtered CommonCrawl data in 100 languages. The large version has 24 transformer layers, 1024 hidden units, 16 attention heads, and approximately 550 million parameters. It is optimized for cross-lingual transfer and performs strongly on multilingual understanding tasks.
- mT5: it is a multilingual extension of the T5 model, trained on the mC4 corpus covering 101 languages. It uses an encoder-decoder architecture, framing all tasks in a text-to-text format. The mT5-Base version contains $\approx$ 580 million parameters, while mT5-Large reaches 1.2 billion.

These models are powerful due to their broad language coverage and scale. However, fine-tuning them directly on English legal texts can degrade their multilingual capabilities—a phenomenon often referred to as catastrophic forgetting. The hypothesis is that when the entire model is updated to optimize for English classification, it loses generalizable features critical for cross-lingual transfer, particularly in morphologically or syntactically different languages.

To mitigate this problem, the authors explored parameter-efficient adaptation strategies, which aim to retain the model's general knowledge while adapting it to specific tasks. These techniques have grown in popularity since around 2020 with the rise of large pre-trained models, and they are now standard practice in multilingual and low-resource NLP. The paper evaluates:

- Partial retraining and fine-tuning: the idea is to freeze the bottom N layers and only train the top layers or classifier head. This reduces over-fitting and computational costs and is widely used even in high-resource domains.
- Adapters: they are small trainable modules (often with a bottleneck structure) inserted between layers—typically after the feed-forward block in each transformer layer. Introduced by Houlsby et al. (2019), adapters enable training with less than 5% of the parameters while maintaining strong performance.
- BitFit: it is a minimal adaptation strategy that only updates the bias terms of the model. Despite its simplicity, BitFit often achieves competitive results, especially in low-data regimes.
- LNFIT: it consits in introduceing low-rank matrices into the attention or feedforward layers. LoRA (Hu et al., 2021) enables efficient adaptation without modifying the original weights, which is especially helpful when memory or storage is limited.

These methods vary in effectiveness depending on the number of training examples, label complexity, and target languages. Nonetheless, they offer a practical compromise between full retraining and performance preservation.

Several studies outside the legal NLP domain have encountered similar challenges in zero- and one-shot classification. For instance, work on multilingual named entity recognition (e.g., CROSSNER by Rahimi et al. [2021]) and sentiment classification (e.g., Amazon Reviews in Keung et al. [2020]) shows that models fine-tuned on a single language often perform poorly when transferred to others. These issues arise from language-specific overfitting and loss of generalizable multilingual representations—problems also highlighted in the MultiEURLEX study when models were retrained solely on English data.

While our state-of-the-art reference remains the MultiEURLEX paper, we found it useful to briefly mention a few related strategies explored in the broader literature, as we may draw inspiration from them in the final stage of our project to improve performance. These include prompt-based reformulations of classification tasks, embedding textual label descriptions, contrastive learning techniques, and synthetic data augmentation. Although not central to our literature review, such approaches offer potential avenues for experimentation.

3 Experiment proposal

To evaluate the effectiveness of adaptation strategies in zero- and one-shot multilingual legal classification, we design a controlled experiment based on the framework proposed by Chalkidis et al. (2021). To balance diversity and feasibility, we restrict the original multilingual setup to a subset of four evaluation languages, each from a different language family featured in the MultiEURLEX dataset: French (Romance), German (Germanic), Polish (Slavic), and Finnish (Uralic). This choice preserves key linguistic variation while reducing the computational load.

Our experiment aims to reproduce and explore two key observations from the original study:

1° Baseline performance: we begin by evaluating a standard baseline— a model fine-tuned only on English. This comparison serves to confirm the performance gap caused by over-fitting or lack of task-specific adaptation in multilingual legal classification.

2° Adaptation strategy assessment: We will then experiment with one or two of the four adaptation strategies cited, selected based on their reported performance and implementation feasibility. Due to limited training time and computational resources, we will focus on a smaller set of methods to allow for more in-depth testing and analysis within our available time frame.

We opted to use XLM-Roberta (Base) for our experiments due to its strong multilingual capabilities, broad language coverage (including all languages retained in our reduced dataset), and proven performance on classification tasks. Compared to encoder-decoder models like MT5, XLM-R is more efficient for multi- label classification, offering faster training and inference. It also benefits from wide support in popular NLP frameworks, making it easier to integrate and fine-tune. Given limited computational resources, XLM-R strikes an ideal balance between performance, scalability, and ease of implementation.

In addition to R-Precision, which is the primary metric used in the original MultiEURLEX paper and well-suited for multi-label classification, we include micro- and macro-averaged F1 scores to better capture performance across both frequent and rare labels. We also report Label Ranking Average Precision (LRAP), which evaluates the quality of label rankings for each document, offering a more nuanced view of prediction confidence. To assess the efficiency of our models, we track training and inference time, model size, and memory usage, enabling us to later compare models not only in terms of accuracy but also computational cost.

Finally, we will attempt to improve performance or reduce cost by applying our own ideas—such as trying to build a classification from scratch by analysing similar patterns (NER, latain vocabulary, patterns in header) in cleaned and lematized laws. Time permitting, we may draw inspiration from broader literature (e.g., prompt-based methods or label embeddings).

Our objective is not to reproduce every aspect of the original work, but to gain practical insights into which adaptation strategies are most effective under realistic constraints, and whether meaningful improvements can be achieved with minimal computational investment.

4 Data

4.1 Dataset overview

We retrieved the MultiEURLEX dataset from the Hugging Face platform, along with label metadata from the authors' official GitHub repository. The dataset contains approximately 65,000 EU legal documents published between 1958 and 2015, available in 23 official EU languages. Following the original paper, we adopt a temporal split to mitigate concept drift (discussed later):

- Training set: ≈ 50,000 documents (1958–2010)
- Validation set: ≈ 5,000 documents (2010–2012)
- Test set: ≈ 5,000 documents (2012–2015)

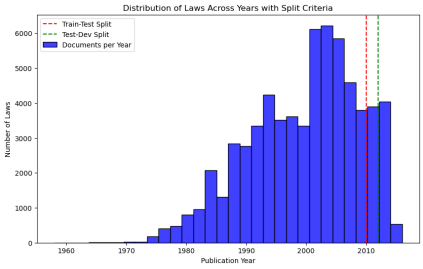


Figure 1 – Distribution of laws

As shown in Figure 1, the volume of published laws has increased sharply over time, peaking between 2005 and 2008.

4.2 Specificities and challenges

4.2.1 Multi-Label and Hierarchical Classification

Each document in the dataset is annotated with multiple labels, reflecting the multi-faceted nature of legal content. These labels are derived from the EuroVoc thesaurus, a hierarchical taxonomy spanning up to eight levels. Label depth varies from one document to another.

To standardize evaluation, the authors expand each label to include its ancestors in the hierarchy. For example, a level-3 label automatically includes its parent labels from levels 1 and 2. At Level 1 (see Table 6 in the Appendix), there are 21 high-level categories such as Politics, International Relations, Law, Economics, Energy, etc. Some categories already seem to partially overlap overlap, potentially introducing ambiguity—e.g., Agriculture and Agri-foodstuffs may cover intersecting topics.

4.2.2 Temporal concept drift

A significant challenge when working with legal data is concept drift—the idea that the meaning, importance, or distribution of topics can shift over time. Legal frameworks evolve, new topics emerge (e.g., digital privacy, green energy), and classification guidelines may change. For this reason, the authors chose a temporal data split rather than random sampling to more realistically simulate a model's ability to generalize to future data.

We follow the same strategy, ensuring that our training set contains only documents published before 2010, while evaluation and test sets contain laws from 2010 onward.

4.3 Language

Although the MultiEURLEX dataset supports 23 official EU languages, not all documents are translated into every language. In total, 12,248 documents lack at least one of the five languages selected for this study (English, French, German, Polish, and Finnish). This is coherent with Table 1 in the original paper.

This is partly explained by the temporal nature of the data split: English, French, and German have consistently been working languages of the European Union and are present in all documents. By

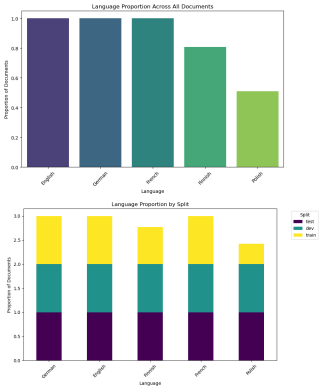


Figure 2 – Language distribution overall and across splits

contrast, languages such as Polish and Finnish were added later—Poland joined the EU in 2004—so many older laws in the training set were never translated into these languages. As a result, only slightly more than half of the training documents include Polish, even though the validation and test sets (covering more recent years) offer full coverage for all selected languages.

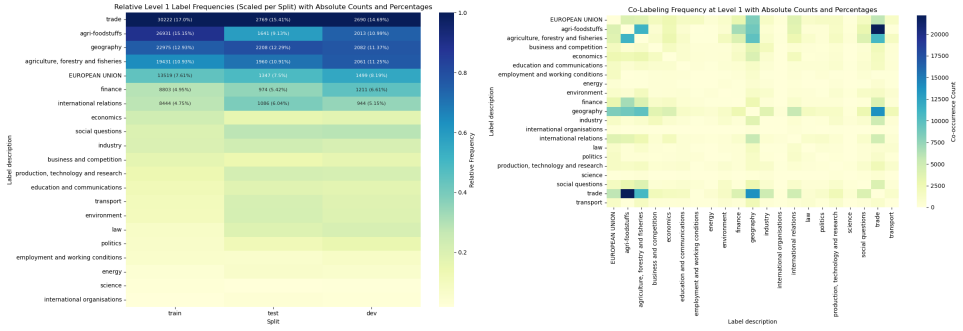
This temporal bias must be considered when evaluating cross-lingual performance, especially on languages with partial historical representation in the training data.

4.4 Label distribution and co-labelling

We focus on Level 1 labels to investigate how label distributions compare across the training, validation, and test sets. An important question is whether certain labels are consistently underrepresented and might therefore be more challenging to predict. Additionally, we examine which labels frequently co-occur within the same documents.

Overall, trade, agriculture-related topics, and geography are the most common labels across all splits, although their relative rankings differ slightly—as illustrated in the heatmap. The largest variations appear among mid-frequency labels: for instance, economics and business are more prevalent in the training set, whereas social questions, industry, and research and technology occur more frequently in the validation and test sets.

These discrepancies can largely be attributed to the temporal data split. Certain types of legislation were more prominent in earlier stages of the EU's development (e.g., economic and business topics), while others (e.g., social and environmental concerns) have gained prominence in more recent years. The similarity between the validation and test distributions further supports this interpretation.



(a) Label distribution accross splits

(b) Co-labelling patterns at level 1

Regarding co-labelling patterns, frequent combinations include labels related to agriculture, trade, and geography. These often serve as broad, general-purpose categories, which helps explain their repeated co-occurrence. Notably, trade and geography appear to function as generic tags that are commonly assigned alongside a wide range of more specific topics.

Nevertheless, this co-labelling analysis is limited to label pairs, while in practice, legal documents are often annotated with multiple (more than two) labels. A more comprehensive analysis of higher-order co-occurrence patterns—such as triplets or clusters of labels—could offer deeper insights into thematic groupings within legal texts. However, due to time constraints, this was left for future work.

4.5 Document length

The legal documents in MultiEURLEX are generally quite long, which is expected given the nature of the content. Most documents fall within the 1,000 to 10,000 token range, with a distribution skewed toward shorter texts but exhibiting a heavy right tail. A small number of documents are extremely long, reaching up to one mil-

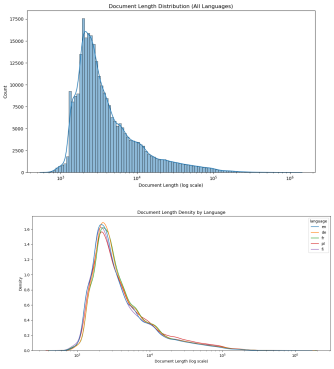


Figure 4 – Document length distribution overall and across languages

lion tokens, which could present challenges for transformer-based models with fixed input lengths.

To better understand these outliers, we examined the ten longest documents in the dataset: six appear in the training set, four appear in the test set and none appear in the development set. Interestingly, many of the longest documents are from the 1993–2002 period. This suggests that certain types of legislation—perhaps foundational treaties, comprehensive regulations, or technical annexes—may have been more common during this era.

The average document length also shows non-uniform variation over time. While there are spikes caused by individual outliers, even the typical document size fluctuates, often ranging between 2,500 and 10,000 tokens. There is a noticeable decrease in average length in the development set compared to training and test sets, which may reflect differences in document type or administrative streamlining in more recent years.

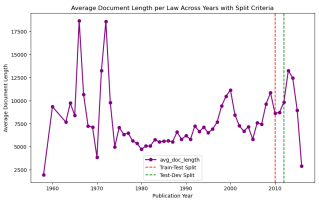


Figure 5 – Document length distribution as a function of year

These patterns may impact how models handle input truncation or long-context reasoning.

5 Results reproduction in simplified setting

Implementation constraints and experimental simplifications

Our reproduction was constrained by several practical challenges that reflect common real-world limitations in legal NLP applications. These included package conflicts, fallback to CPU-only training, and limited storage and compute availability. As a result, we deviated from the original MultiEURLEX setup in several key ways: we limit training to a subset of 5,000 to 10,000 examples, use larger batch sizes (32) to accelerate iteration, and reduce the number of epochs to 2 instead of 5. In addition, we restrict classification to the Level 1 EuroVoc hierarchy (21 labels), instead of Level 3 (over 500 labels), in order to reduce output dimensionality and training complexity.
While these changes inevitably lowered absolute performance, they allowed us to evaluate whether the same qualitative trends and conclusions still hold—particularly regarding adaptation strategies like freezing and adapters.

5.1 First result reproduction: Performance drop from English-only fine-tuning

To establish a baseline, we retrained XLM-Roberta (Base) solely on English-language training data, then evaluated its performance across five target languages on Level 1 labels.

5.1.1 Training performance and costs

The training process confirmed the high cost of fine-tuning transformer models, both in time and memory. Despite convergence (low loss and decent AUC), the model overfit to English syntax and semantics, leading to poor generalization to other languages—a symptom of catastrophic forgetting in multilingual settings.

Model	Training time	Memory used	Loss	AUC
XLM-Roberta	18675.41	15473.85 MB	0.3483	0.55

Table 1 – Training performance of the model

5.1.2 Multilingual evaluation performance

Evaluation across languages confirms that fine-tuning XLM-Roberta exclusively on English data leads to poor multilingual generalization, with consistently low performance across all tested languages. Micro F1 scores remain clustered between 0.2477 (Finnish) and 0.2639 (German), while Macro F1 hovers around 0.05, indicating that the model struggles to handle both common and rare labels. Notably, no language significantly outperforms English, and the small differences observed—such as German slightly outperforming French and Polish—are too minor to indicate robust cross-lingual generalization. LRAP scores, which range from 0.4779 to 0.4931, are somewhat more stable, suggesting the model retains some capacity to rank relevant labels, even if its discrete predictions are weak.

Language	R-mean precision	Micro F1	Macro F1	LRAP
English	0.2658	0.2544	0.0499	0.4863
Finnish	0.2663	0.2477	0.0489	0.4779
French	0.2681	0.2622	0.0509	0.4928
German	0.2651	0.2639	0.0514	0.4931
Polish	0.2668	0.2564	0.0506	0.4844

Table 2 – Performance of multi-lingual model retrained in English

A likely explanation for these uniformly low results is insufficient training (c.f. Implementation constraints). Under such limited conditions, the model may not even reach optimal performance on English, preventing clearer performance gaps across languages. In more extensive training regimes, we would expect English to perform significantly better than the others, and for patterns of linguistic similarity to become more evident—e.g., French and German outperforming Polish and Finnish. Additionally, the use of AUC as the training objective, while helpful for label ranking, likely contributed to poor Micro and Macro F1, as it does not directly optimize for classification accuracy. This mismatch between objective and evaluation may further explain the relatively better LRAP scores compared to thresholded F1 metrics. Future work could benefit from more aligned loss functions or post-training calibration to improve classification outcomes.

5.1.3 Catastrophic forgetting and experiment proposal

This performance drop is a known phenomenon in multilingual NLP and is often referred to as catastrophic forgetting—the model loses previously acquired cross-lingual knowledge when heavily adapted to one language. This reinforces the need for more efficient adaptation strategies that preserve multilingual capabilities.

Due to time constraints, we did not pursue a deeper investigation, but we outline a proposed analysis and include a draft implementation in the accompanying notebook. The goal would be to better understand how fine-tuning induces language-specific degradation:

- Save XLM-R model weights before and after fine-tuning .
- Compute L2 norm differences layer by layer to identify where weight shifts are most pronounced
- Visualize change patterns to determine whether early (language-general) or later (task-specific) layers are more affected—insights that could guide future partial fine-tuning or freezing strategies.

5.2 Second result reproduction: "better" performance with adaptation strategies

5.2.1 Forzen layers

This experiment evaluates the effect of freezing the first N encoder layers of XLM-R during English-only training. The goal is to balance multilingual knowledge retention and task-specific adaptability. While absolute metric differences may appear small, general patterns reveal meaningful trade-offs:

Low Freezing (N = 3): With minimal constraint, the model exhibits moderate, language-balanced performance (e.g., Micro F1 $\approx$ 0.31 for English, German), but at high memory (15.3 GB) and training time costs. AUC $\approx$ 0.5 suggests under-fitting, likely due to limited training (2 epochs) and small data size.

Moderate Freezing (N = 6): This setting achieves the best multilingual balance. Polish and German performance improves significantly (Micro F1 = 0.3680 and 0.3348, respectively), indicating effective task adaptation while preserving cross-lingual capacity. Training is more efficient, and the slightly lower AUC remains consistent with label sparsity.

High Freezing (N = 9): As more layers are locked, adaptability declines. Metrics regress to near N = 3 levels across languages, confirming under-fitting due to insufficient parameter flexibility.

Full Freezing (N = 12): Surprisingly, English performance jumps (Micro F1 = 0.6334), with moderate gains in other languages. Since the model is not fine-tuned at all here (all layers frozen), this suggests the strong alignment of pre-trained English features with the classification task. However, this comes at the cost of true multilingual generalization, especially for Polish (Micro F1

= 0.3074). The spike in AUC (0.6321) likely reflects English-specific overconfidence rather than improved learning.

Overall, N = 6 emerges as the optimal compromise, confirming findings from the MultiEURLEX paper: moderate adaptation preserves multilingual capabilities while boosting performance in low-resource or typologically distant languages. These results also underscore a known limitation of multilingual models — they are strongest in English and require careful tuning to generalize elsewhere.

# frozen layers	Language	R-mean precision	Micro F1	Macro F1	LRAP
N = 3 Training - 5223s, Memory - 15348 MB AUC - 0.5007	English	0.2722	0.3122	0.0559	0.5475
	Finnish	0.2668	0.2938	0.0544	0.5277
	French	0.2830	0.3017	0.0564	0.5432
	German	0.2721	0.3116	0.0575	0.5410
	Polish	0.2709	0.2860	0.0537	0.5138
N = 6 Training - 7139s, Memory - 12116 MB AUC - 0.4999	English				
	Finnish	0.2745	0.3222	0.0579	0.5582
	French	0.28	0.3176	0.0578	0.5469
	German	0.2744	0.3348	0.0597	0.5538
	Polish	0.2767	0.3680	0.0634	0.5823
N = 9 Training - 4935s, Memory - 11548 MB AUC - 0.5006	English	0.2722	0.3122	0.0559	0.5517
	Finnish	0.2668	0.2938	0.0544	0.5321
	French	0.2830	0.3017	0.0564	0.5449
	German	0.2721	0.3116	0.0575	0.5446
	Polish	0.2709	0.2860	0.0537	0.5208
N = 12 Training - 4400s, Memory - 7147 MB AUC - 0.6321	English	0.3128	0.6334	0.3595	0.7746
	Finnish	0.2829	0.4222	0.1179	0.6282
	French	0.3072	0.3841	0.1408	0.6186
	German	0.2955	0.4497	0.1631	0.6542
	Polish	0.2873	0.3074	0.1237	0.6316

Table 3 – Performance of multi-lingual model retrained in English with N first layers frozen (out of 12 layers)

5.2.2 Adapters

This final experiment investigates the effectiveness of using adapter layers as an alternative to full fine-tuning or layer freezing. Adapters are lightweight, trainable modules inserted within each transformer block that allow task-specific adaptation without updating the full model weights.

Despite our implementation constraints the adapter-based model achieves the **strongest overall multilingual performance** across all evaluated strategies. Micro F1 scores range from 0.4238 (Polish) to 0.4511 (German), substantially outperforming both the English-only baseline and the frozen-layer configurations (even the optimal N=6 setup). The improvements extend to Macro F1 and LRAP, which also reach their highest levels in this experiment, suggesting better handling of rare labels and more accurate label ranking. For instance, LRAP exceeds 0.65 in English and German, highlighting strong probabilistic predictions even in multi-label settings.

Model	Language	R-mean precision	Micro F1	Macro F1	LRAP
Adapters Training - 2486s, Memory - -167 MB AUC - 0.5266	English	0.2905	0.4477	0.1525	0.6519
	Finnish	0.2922	0.4409	0.1389	0.6331
	French	0.2888	0.4470	0.1265	0.6357
	German	0.2873	0.4511	0.1413	0.6569
	Polish	0.2884	0.4238	0.1401	0.6204

Table 4 – Performance of multi-lingual model retrained in English with adapters

Adapters also offer practical advantages: training time is relatively low (2486s, faster than full fine-tuning), and memory use is reduced compared to updating the full model. This demonstrates that **adapter-based training is not only performance-effective but also resource-efficient**.

Nevertheless, it is likely that this setting under-represents the full potential of adapters. With access

to more data, training epochs, or a deeper label hierarchy, performance would almost certainly increase, as suggested by results from the original MultiEURLEX study.

Finally, while we explored individual strategies in isolation, hybrid approaches—such as combining frozen lower layers with adapters in higher layers—may offer further gains in both generalization and efficiency. Unfortunately, due to resource constraints, such configurations remain outside the scope of this reproduction. This highlights a broader challenge in real-world NLP research: the trade-off between experimental thoroughness and practical feasibility.

6 How to get better performance or lower cost?

6.1 Domain-specific lexical cues for legal texts

Beyond adaptation-based fine-tuning, we propose a lightweight, exploratory method leveraging domain-specific lexical patterns in legal texts. Legal documents often share structural and lexical features—such as Latin-root terms, legal formulas, and references to directives—that may generalize across languages.

The idea is to statistically link tokens (or lemmas) to labels and assess their consistency across languages. For instance, consistent co-occurrence of environment-related terms in both English and Polish could support cross-lingual generalization or inform efficient embeddings.

Though not fully implemented due to time constraints, our accompanying notebook outlines the approach from preprocessing and lemmatizing texts using spaCy to computing token–label co-occurrence statistics and comparing cross-lingual patterns to identify shared lexical signals.

This interpretable strategy could come as a complement to large models by revealing stable domain cues.

6.2 Label-aware adaptation: prompting and embedding-based matching

To complement token-level analyses and fine-tuning strategies—especially under compute and time constraints—we explored lighter, label-aware alternatives: prompt-based reformulations and embedding-based label matching. These methods directly leverage the semantic knowledge of pre-trained multilingual encoders, avoiding the need for task-specific training or classifier heads.

While their absolute performance is limited, they offer valuable trade-offs: significantly reduced training cost, faster deployment, and broader cross-lingual generalization. By aligning documents with label semantics either through natural language prompts or shared embedding spaces, these strategies present strong, scalable baselines for one-shot legal classification in multilingual and low-resource scenarios.

6.2.1 Prompt based reformulation

We explored prompt-based reformulation, which frames multi-label classification as a natural language inference task. Instead of training a classification head, labels are embedded directly into input prompts using natural language templates. This allows the model to leverage its pre-trained multilingual understanding to associate documents with legal categories—without any task-specific updates to the model weights. Such an approach is particularly suited to zero-shot and cross-lingual scenarios.

Two prompt styles were tested:

- generic prompt: “This legal document discusses the following topics: text. What legal categories apply?”
- guided prompt: “This legal document is about: text. Possible categories include: politics, economics, environment, health, ... Which ones apply?”

While the guided prompt provided slightly clearer task alignment, both versions led to minimal improvements over the baseline. Performance metrics such as Macro F1 remained low ($\approx$ 0.054–0.057), indicating limited ability to detect rare classes.These marginal improvements came at a relatively high computational cost—$\sim$ 15 GB memory and $\sim$ 1.5 hours of proessing time. We

attribute the modest results to a combination of factors: the simplicity of the prompts, the limited legal-domain alignment of the pre-trained model, and the absence of task-specific adaptation. Despite the theoretical promise of prompt-based methods, especially in zero-shot settings, our results suggest that more sophisticated prompt engineering or hybrid approaches may be necessary to realize their full potential in legal NLP tasks.

Original adaptation strategy	Language	R-mean precision	Micro F1	Macro F1	LRAP
prompt-based training Training - 5567s, Memory - 15451 MB AUC - 0.50	English	0.2722	0.3122	0.0559	0.5505
	Finnish	0.2668	0.2938	0.0544	0.5309
	French	0.2830	0.3017	0.0564	0.5447
	German	0.2721	0.3116	0.0575	0.5433
	Polish	0.2709	0.2860	0.0537	0.5180
embedding-based label matching Training - None	English	-	0.2631	0.1001	0.2179
	Finnish	-	0.2654	0.0938	0.2176
	French	-	0.2649	0.0941	0.2174
	German	-	0.2637	0.0923	0.2170
	Polish	-	0.2634	0.0907	0.2167

Table 5 – Performance of alternative strategies

6.2.2 Label embedding matching

This approach reframes classification as a similarity task in embedding space. Instead of training a model, we compute dense vector representations for both documents and label descriptions using a pre-trained multilingual encoder, and select the top-5 labels based on cosine similarity. This enables one-shot classification, where a single textual description per label suffices—eliminating the need for task-specific fine-tuning.

Despite its simplicity, the method demonstrated promising results in handling rare or nuanced legal categories. While overall performance remained modest (Micro F1: 0.26, LRAP: 0.22), Macro F1 doubled compared to the prompt-based baseline ($\sim$ 0.09 vs. $\sim$ 0.045), highlighting improved sensitivity to infrequent labels—a critical aspect of imbalanced legal classification.

Though it lacks decision boundaries learned through supervision, this training-free, fast, and scalable method is especially valuable in low-resource or exploratory multilingual settings. Its performance could likely be improved through better alignment between label descriptions and model pretraining, or through hybrid approaches involving lightweight adaptation.

7 Conclusion

This report explored the challenges and opportunities of multilingual and zero-shot legal text classification in NLP, focusing on the MultiEURLEX dataset and the framework proposed by Chalkidis et al. We emphasized the difficulties of relying on large, labeled datasets, particularly for low-resource languages, and examined several adaptation strategies—including freezing layers and using adapters—to address the issue of catastrophic forgetting in cross-lingual transfer. Our experiments, though constrained by limited computational resources, confirmed that moderate adaptation (e.g., freezing 6 layers) and parameter-efficient methods like adapters offer strong performance trade-offs. Adapters, in particular, demonstrated superior multilingual generalization while remaining resource-efficient. Additionally, we proposed an exploratory, language-agnostic approach based on lexical cues and token-level patterns to complement large model training. Overall, this work underscores the importance of efficient model adaptation and domain-specific insight in building scalable, inclusive legal NLP systems. To better capture these trade-offs, we also expanded the evaluation beyond standard performance metrics to include measures of model cost—such as training time and model size—as well as a broader set of quality metrics.

References

Ilias Chalkidis, Manos Fergadiotis, Ion Androutsopoulos, and Nikolaos Aletras. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6970–6984. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.553.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2020. URL https://arxiv.org/abs/1911.02116.

Afshin Rahimi, Zheng Yuan, and Trevor Cohn. Crossner: Evaluating cross-lingual named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 624–634, 2021.

Phillip Keung, Yichao Lu, and Siddharth Goyal. Multilingual representation learning via machine translation fine-tuning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 153–162, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mohammad Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Appendix

Metrics used

AUC - area under the Receiver Operating Characteristic (ROC) curve:

AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Formally, it is the integral of the True Positive Rate (TPR) over the False Positive Rate (FPR) across all possible classification thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(x)\,dx, \quad \text{where} \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

This threshold-independent nature makes AUC especially informative in imbalanced multi-label settings where the choice of decision threshold can significantly bias evaluation. While it does not directly reflect classification accuracy, AUC remains useful in diagnosing whether the model is generally capable of distinguishing relevant from irrelevant labels before thresholding is applied.

LRAP - Label Ranking Average Precision:

LRAP evaluates how well the model ranks the true labels ahead of irrelevant ones for each instance. Specifically, it computes, for each true label, the fraction of true labels ranked above it, averaged across all true labels.

$$\text{LRAP}_i = \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' \in Y_i : \text{rank}_i(y') \leq \text{rank}_i(y)\}|}{\text{rank}_i(y)}, \quad \text{where} \quad \text{LRAP} = \frac{1}{n} \sum_{i=1}^{n} \text{LRAP}_i$$

Unlike AUC, LRAP is label-sensitive and rewards models that not only separate classes but also order them correctly. This makes it particularly suited for applications like legal tagging, where multiple labels may be correct and ranked output is often more actionable than hard classification.

Macro F1

Computes the F1 score independently for each label and then averages across labels, assigning equal weight to all, regardless of frequency.

$$\text{F1}_l = 2 \cdot \frac{\text{Precision}_l \cdot \text{Recall}_l}{\text{Precision}_l + \text{Recall}_l}, \quad \text{F1}_{\text{macro}} = \frac{1}{L} \sum_{l=1}^{L} \text{F1}_l$$

This highlights whether the model is able to generalize beyond the most common legal categories. In the highly imbalanced EuroVoc label space, where many concepts are under-represented, Macro F1 provides a critical lens into model fairness and robustness.

Micro F1:

Balances between precision and recall while aggregating all true positives, false positives, and false negatives globally across all labels, making it heavily influenced by the performance on frequent labels. This is useful for understanding overall classification capacity but tends to obscure the model's ability to detect rare or niche concepts.

$$\text{Precision}_{\text{micro}} = \frac{\sum_l \text{TP}_l}{\sum_l (\text{TP}_l + \text{FP}_l)}, \quad \text{Recall}_{\text{micro}} = \frac{\sum_l \text{TP}_l}{\sum_l (\text{TP}_l + \text{FN}_l)}, \quad \text{F1}_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}$$

R-mean precision:

It is a document-level metric that evaluates whether the model retrieves the correct number of labels per instance. For each document, R-Precision is computed as the proportion of correct labels among

the top R predictions, where R is the number of ground truth labels for that document. The mean is then taken over all test samples.

$$\text{R-Precision}_i = \frac{|\hat{Y}_i \cap Y_i|}{R_i}, \quad \text{Mean R-Precision} = \frac{1}{n}\sum_{i=1}^{n} \text{R-Precision}_i$$

This metric effectively balances precision and recall in a controlled, per-document setting, and is particularly appropriate when evaluating predictions in multi-label classification scenarios, where the number of correct labels varies widely between samples.

Eurovoc labels

Level	Values
level_1	21
level_2	127
level_3	567
level_4	3861
level_5	2284
level_6	481
level_7	43
level_8	6

(a) 8 labeling levels of the Eurovoc law classification system

Level 1 labels	label description
100142	politics
100143	international relations
100144	EUROPEAN UNION
100145	law
100146	economics
100147	trade
100148	finance
100149	social questions
100150	education and communications
100151	science
100152	business and competition
100153	employment and working conditions
100154	transport
100155	environment
100156	agriculture, forestry and fisheries
100157	agri-foodstuffs
100158	production, technology and research
100159	energy
100160	industry
100161	geography
100162	international organisations

(b) Level 1 labels descriptions

Table 6 – Eurovoc label repartition and description

Comparison with article

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).

Figure 6 – Statistics per languages

	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.

Figure 7 – Tests results