MEL: LEGAL SPANISH LANGUAGE MODEL

David Betancur Sánchez
Instituto de Ingeniería del Conocimiento (IIC)
david.betancur@iic.uam.es

Nuria Aldama García
Instituto de Ingeniería del Conocimiento (IIC)
nuria.aldama@iic.uam.es

Álvaro Barbero Jiménez
Instituto de Ingeniería del Conocimiento (IIC)
alvaro.barbero@iic.uam.es

Marta Guerrero Nieto
Instituto de Ingeniería del Conocimiento (IIC)
marta.guerrero@iic.uam.es

Patricia Marsà Morales
Instituto de Ingeniería del Conocimiento (IIC)
patricia.marsa@iic.uam.es

Nicolás Serrano Salas
Instituto de Ingeniería del Conocimiento (IIC)
nicolas.serrano@iic.uam.es

Carlos García Hernán
Instituto de Ingeniería del Conocimiento (IIC)
carlos.garcia@iic.uam.es

Pablo Haya Coll
Instituto de Ingeniería del Conocimiento (IIC)
pablo.haya@iic.uam.es

Elena Montiel Ponsoda
Ontology Engineering Group
Universidad Politécnica de Madrid
elena.montiel@upm.es

Pablo Calleja Ibáñez
Ontology Engineering Group
Universidad Politécnica de Madrid
pcalleja@fi.upm.es

ABSTRACT

Legal texts, characterized by complex and specialized terminology, present a significant challenge for Language Models. Adding an underrepresented language, such as Spanish, to the mix makes it even more challenging. While pre-trained models like XLM-RoBERTa have shown capabilities in handling multilingual corpora, their performance on domain specific documents remains underexplored. This paper presents the development and evaluation of MEL, a legal language model based on XLM-RoBERTa-large, fine-tuned on legal documents such as BOE (Boletín Oficial del Estado, the Spanish oficial report of laws) and congress texts. We detail the data collection, processing, training, and evaluation processes. Evaluation benchmarks show a significant improvement over baseline models in understanding the legal Spanish language. We also present case studies demonstrating the model's application to new legal texts, highlighting its potential to perform top results over different NLP tasks.

*K*eywords Natural language processing · Information extraction · Spanish legal domain · Language Model

1 Introduction

Background. The legal domain presents challenges for Natural Language Processing (NLP) solutions due to the highly specialized nature of legal texts, specific terminology, precise references to laws, and archaic language that often differs from everyday communication. Recent advances in NLP, particularly since the introduction of the transformer architecture [1] and transformer-based models such as BERT [2] and XLM-RoBERTa [3], have demonstrated remarkable abilities to understand human language in a wide range of domains and languages, including Spanish [4]. However,

while these models perform well on general text, their effectiveness in a highly specific domain such as the legal domain, especially those written in Spanish, is still underexplored. Some multilingual legal corpora like the MultiLegalPile [5] are available but, as it's multilingual, the focus on specific languages is not usually guaranteed and can fall down on size or even quality of cleaning. Proprietary models like legal-xlm-roberta-large [6] use this corpus but still achieve similar results to the base model (xlm-roberta-large) at Spanish tasks evaluated in our benchmark at section 4.2.

Motivation. In recent years, the application of NLP models to domain-specific tasks has gained increasing attention, particularly in areas such as law and medicine. Legal professionals frequently work with large volumes of complex documents and the manual analysis of these documents is time consuming, prone to human error, and often requires expert knowledge. As a result, there is a growing demand for automated tools that can assist in processing and extracting meaningful information from these legal texts. However, it is important to note that Spanish is usually underrepresented in the training corpora of large language models, creating a challenge for developing NLP tools for Spanish legal texts.

Objective. This paper aims to present MEL (Modelo de Español Legal), a legal language model based on XLM-RoBERTa-large for understanding Spanish legal texts. The model will be trained on a curated dataset of Spanish legal documents and evaluated against baseline models to measure its performance in various NLP tasks, such as legal classification and named entity recognition (NER). The goal is to demonstrate the model's ability to outperform existing multilingual models in accurately capturing the features of legal Spanish, offering a valuable tool for legal professionals and NLP researchers working on these domain specific tasks.

Content. In this paper, first we discuss related work on domain-specific language models, multilingual and underrepresented languages, and legal Spanish models and their challenges. After that we will go through the methodology used to obtain the training corpus, train and benchmark our model. Finally, the results of both the training and benchmark against another models.

2   Related Work

Encoder-only vs LLMs. Despite the recent hype with generative LLMs, encoder-only models remain widely used in a variety of non-generative downstream applications [7]. They offer a great trade-off between quality and size, making it's inference requirements much better than the ones offered in LLMs. Also, on tasks such as NER and information extraction, generative models still lack intuition about correct positioning the entities, shown on the experiments performed on [8].

Domain-Specific Language Models. The adaptation of pre-trained language models for specific domains has become a prominent area of research in NLP. Models like BioBERT [9] and SciBERT [10], fine-tuned for biomedical and scientific texts, respectively, have shown significant performance improvements over general-purpose models in domain-specific tasks. Legal NLP has followed a similar trajectory, with models like Legal-BERT [11] and CaseLawBERT [12] fine-tuned on legal corpora, showing superior performance in tasks such as contract analysis, court judgment summarization, and legal question answering. These works underscore the importance of domain-specific fine-tuning for achieving state-of-the-art results.

Multilingual and open-access data underrepresented Languages. Multilingual models like XLM-RoBERTa and mBERT have significantly improved text understanding across multiple languages, including underrepresented ones like Spanish. In some cases, they even outperform monolingual models, particularly because many existing monolingual models for Spanish are older. However, their effectiveness can still vary depending on the specific domain and task at hand. [13]

Legal NLP Spanish Models. Recent work combines both domain adaptation and multilinguality. A notable example is legal-xlm-roberta-large [14], a multilingual transformer pre-trained on the Multi-Legal-Pile [15], a comprehensive multilingual dataset derived from diverse legal sources across 24 languages. This pre-training approach has demonstrated promising results in legal document classification and NER tasks, showcasing the benefits of incorporating domain-specific and multilingual corpora. However, despite these advancements, the application of such models to Spanish legal texts, specifically using datasets like BOE and parliamentary records, remains underexplored.

Challenges in Legal Language. Despite the above advancements, there remains a gap in models trained specifically on Spanish legal texts. These texts often contain archaic and highly specialized terms and structures that differ significantly from everyday language. Additionally, legal texts require precise interpretation, as minor differences in documents can alter the meaning of a law or legal ruling. This complexity, added to the particularity of underrepresented

languages like Spanish, and the errors presented on OCR transcription, constitute a new layer of difficulty when processing legal texts. Our research seeks to address these challenges by pre-training a multilingual transformer model on Spanish legal corpora, enabling better understanding and application of legal language in NLP tasks such as text classification and NER.

3 Methodology

3.1 Corpus collection

To pre-train MEL, we curated a specialized corpus consisting of various legal documents, including the Boletín Oficial del Estado (BOE), parliamentary proceedings, legal statutes and court rulings [16]. These documents were collected from publicly available Spanish repositories and legal databases.

The data extracted from the official journals, referred to as Boletines Oficiales, were sourced at the notice and rule level wherever possible. These journals were retrieved as PDF files from the websites of the Spanish State and its 20 autonomous communities. Given the diverse formats and structures of these sources, 20 specific web trackers were developed, one for each autonomous community. In certain cases, it was not possible to extract individual notices or rules directly as PDFs; for example, in the autonomous city of Ceuta, entire journal PDFs had to be downloaded, or the transcriptions available on the bulletin websites were used instead. Furthermore, in bilingual autonomous communities, the bulletins were downloaded considering language distinctions. It was observed that some notices were published in one language, with only the titles translated into the other, necessitating careful handling to prevent introducing unintended multilingual data into the corpus.

The PDF files adhere to Directive (EU) 2016/2102 [17], which ensures accessible protocols, facilitating accurate transcriptions. To ensure the extraction of relevant information, extraneous elements such as margins and stamps were cropped, isolating the content related to notices and laws. Furthermore, a comprehensive evaluation of various transcription tools and Optical Character Recognition (OCR) systems was conducted. The tools compared included OCRmyPDF, pdftotext, PDFPlumber and PyPDF. Following this analisys, PDFPlumber was identified as the optimal solution for this specific use case [18].

3.2 Corpus processing

The corpus underwent a preprocessing state to standardize and clean the data. This process involved:

- **Language Cleaning**: We ensured that all documents were written in Spanish, removing any non-Spanish content to maintain linguistic consistency throughout the corpus. This was performed using Facebooks fasttext library [19] selecting the texts that were mainly in Spanish with a confidence higher than 95%.

- **Unwanted characters Cleaning**: The text was cleaned to remove unwanted characters such as extra whitespaces and breaklines. This helped to standardize the corpus and minimize noise during model training.

- **Chunking**: Current model architectures have a predefined context window. In our case the window was of 512 tokens. Texts were split in sentences and appended until reaching a maximum of 512 tokens to avoid unnecessary truncation in the training.

- **Size**: The final corpus after tokenization and chunking consists of 5 520 000 texts, resulting on a dataset of 92.7 GB.

3.3 Model training

A domain adaptation was performed using the Huggingface Transformers Trainer [20]. The following are some important aspects relevant to the training of the model:

- **Training setup and architecture**: The base model used for this task was XLM-RoBERTa-large, a transformer model pre-trained on multilingual data. Different studies such as [13] demonstrate that XLM-RoBERTa-large is the best model across the encoder-only available, scoring even better than the much newer mDeBERTa.

  In this phase, we extended the pre-training using a large corpus on Spanish legal texts, rather than finetuning the model for specific tasks. The architecture remained consistent with XLM-RoBERTa-large, using the same transformer layers, while the training objective focused on masked language modeling (MLM) to reinforce context-based learning in the legal domain. The whole word mask strategy was maintained with random selection of whole words replacing 80% by [MASK], 10% by a random token and 10% left unchanged.

- **Training Data**: The dataset for the continuation of pre-training included a vast collection of Spanish legal texts from BOE, parliamentary transcripts, court, rulings and other legislative documents. Processing of the data is mentioned in the section above. A random selection of 100 000 texts was used for validation of the training.
- **Hyperparameters**:
  - **learning rate**: 1×10^{-4}, selected after the one used in the base model.
  - **Optimizer**: The AdamW optimizer was used, with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1 \times 10^{-6}$. These values were selected after the good performance on the base model.
  - **Batch size and gradient accumulation**: Batch size of 16 with a gradient accumulation of 4. Effective batch size of 64, helping the model process large chunks of data while keeping memory usage manageable. This is the max batch size that fits in the Hardware used, an NVIDIA A100 with a memory of 80GB.
  - **Epochs**: 2
  - **Warmup**: 0.08. Meaning 8 % of the total steps were used to gradually increase the learning rate, as seen on Figure 3, helping the model prevent from making abrupt weight updates in early training.
  - **Weight decay**: 0.01
  - **Scheduler**: A cosine learning rate scheduler was used. Shown on Figure 3.
- **Hardware**: Given the scale of the pre-training (large batch sizes and complexity of large models), high performance hardware with substantial GPU memory and processing power is required. In this case a NVIDIA A100 80GB PCIe was used to complete the training.
- **Compute time**: The continuation of the pre-training took approximately 13.9 days with each epoch lasting between 6 and 7 days. Early stopping was not applied since the objective was to pretrain the model over a fixed number of epochs.

3.4 Evaluation

The evaluation of MEL was challenging because suitable legal text datasets for machine learning tasks are notoriously scarce, particularly for non-English languages such as Spanish. The lack of labeled legal data is a significant barrier to the development of robust fine-tuned models. Legal documents often require specialized annotation by domain expert linguists, making the labeling process both time-consuming and costly. For MEL, two benchmarks were performed both on a public and a private dataset:

- **Multieurlex Spanish multilabel dataset** [21]: This dataset comprises 65000 EU laws annotated with EUROVOC concepts by the EU Publication Office. Each law is associated with one or more EUROVOC labels, where each label corresponds to a specific legal topic. For example, labels might include descriptors such as [60, agri-foodstuffs], [6006, plant product], [1115, fruit]. This descriptors, and their associated IDs, are available in 23 official EU languages, including Spanish.
- **Multiclass classification on private dataset**: This task was conducted on a private legal corpus with texts labeled on different classes. The corpus contains a train set with 2389 texts and a test with 892 texts. There is a total of 9 different classes for the model to predict on.

The models used for the benchmark to compare against MEL include:

- **xlm-roberta-large** [3]: FacebookAI's multilingual version of RoBERTa. It is pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages.
- **PlanTL-GOB-ES/RoBERTalex** [22]: PlanTL (Plan de impulso de las Tecnologías del Lenguaje) transformer-based masked language model for the Spanish language. It is based on the RoBERTa base model and has been pre-trained using a large Spanish Legal Domain Corpora, with a total of 8.9 GB of text.
- **joelniklaus/legal-xlm-roberta-large** [6]: Multilingual model pretrained on legal data. It is based on XLM-R large. For pre-training, Multi Legal Pile [15] was used, a multilingual dataset from various legal sources covering 24 languages.

4 Results

4.1 Training Performance

The training phase demonstrated consistent improvements in the training steps, as shown in Figures 2 and 1. The training loss shows a steady decline, indicating that the model effectively learned to fit the training data. However,

the evaluation loss decreased also, reflecting the model's capacity to generalize and adapt to the features of the legal Spanish texts. Some aspects of the training results include:

- **First steps of training**: During the initial steps of the training, both train and eval loss showed rapid declines, indicating that the model quickly captured fundamental patterns in the data. Warmup steps, used as in the base model training strategy, seemed to allow gradual adaptation, preserving the useful representations learned during pretrained while allowing fine-tuning for the new domain.
- **Final steps of training**: In the later stages, the losses converged more gradually guaranteeing stability of the training.
- **Overfitting check**: The similar steep or parallel trend of the training and evaluation loss throughout training suggests minimal overfitting, further emphasizing the generalizability of the model.
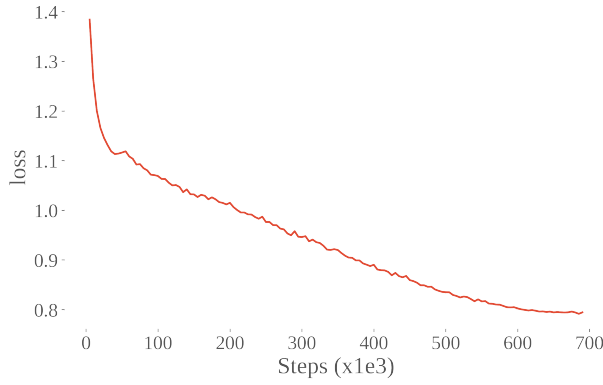

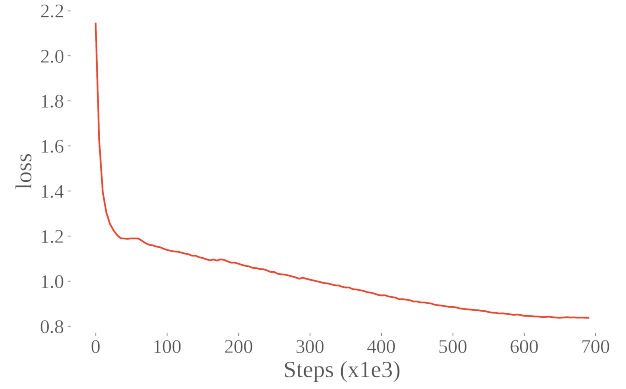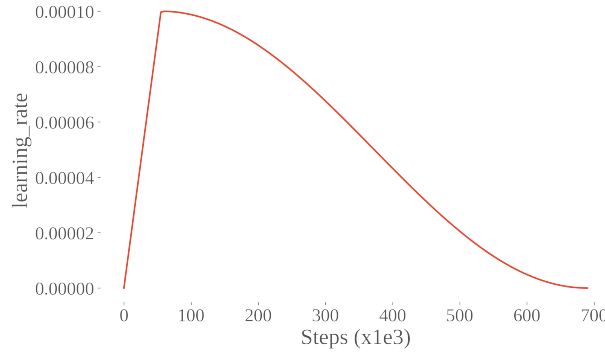Figure 1: Evaluation Loss

Figure 2: Training Loss

Figure 3: Learning rate

4.2   Benchmark Results

As mentioned in section 3.4, to evaluate the pre-trained model, we conducted a comprehensive benchmarking analysis on multilabel and multiclass classification using both multieurlex and a private legal corpus. The F1 score over the finetuning epochs was saved and plotted of Figures 4 and 5. Sometimes, models tend to understand the tasks on similar proportions and end having similar scores, but at different rates. That is why we considered important to evaluate the learning speed of the models represented by the area under the curve of the F1 vs Epochs plot. These results can be visualized on Tables 1 and 2.

- **Multieurlex Spanish multilabel dataset**: On the multieurlex dataset, our model seems to excel, achieving higher F1 scores at a higher learning speed than the rest of the models. It also seems that our model achieves good results on early epochs, meaning that the model can achieve good results with low or nearly-none finetuning.
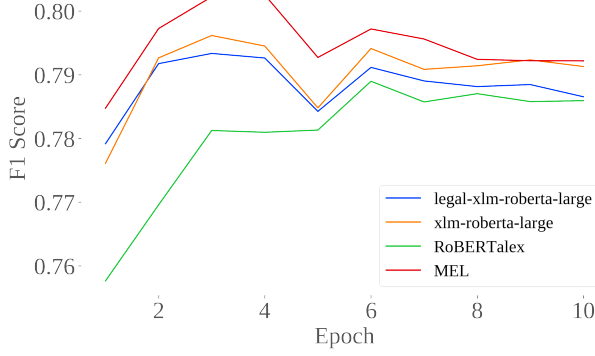
5

Figure 4: Multieurlex F1 vs Epochs

Model	Max F1	Epochs vs F1 AUC
Legal-XLM-RoBERTa-Large	0.7933	7.1016
XLM-RoBERTa-Large	0.7962	7.1205
RoBERTalex	0.7890	7.0324
MEL	**0.8025**	**7.1606**

Table 1: Max F1 scores and learning speed for various models on the Multieurlex dataset. Bold values indicate the best scores.

- **Private Multiclass dataset**: Similar to the evaluation on the eurlex dataset, MEL seems to begin the finetuning with a lot of alredy-known information, achieving higher results after 1 epoch than the RoBERTalex model after a complete finetune. MEL results in higher F1 and learning speed than the rest of the models.
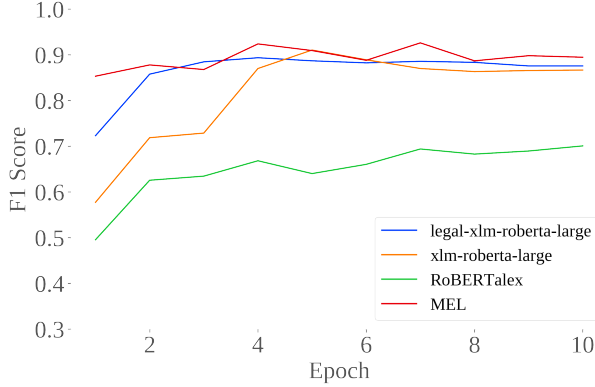

Figure 5: Multiclass F1 vs Epochs

Model	Max F1	Epochs vs F1 AUC
Legal-XLM-RoBERTa-Large	0.8935	7.8487
XLM-RoBERTa-Large	0.9103	7.4372
RoBERTalex	0.7007	5.8929
MEL	**0.9260**	**8.0510**

Table 2: Max F1 scores and learning speed for various models on the Multieurlex dataset. Bold values indicate the best scores.

Following the results on Figure 5, some additional experiments were performed to analyze the behavior of the models on small data. To achieve this, steps previous to the first epoch were saved to explore how the model would perform when training on fewer data. Results can be observed on Figure 6 and Table 3. These results confirm that MEL is better aligned with the legal language, since with small data it is already able to achieve a significantly better F1 than the other models.
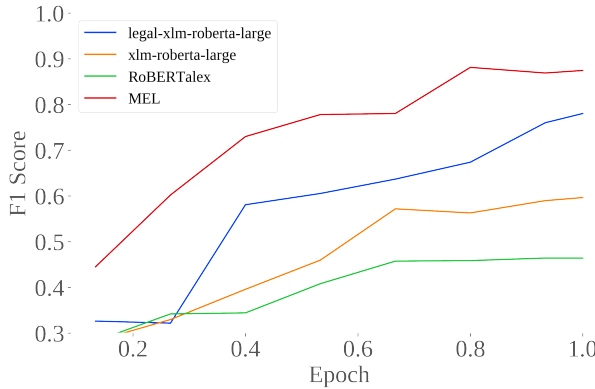

Figure 6: Multiclass F1 vs Epochs (small data)

Model	Max F1	Epochs vs F1 AUC
Legal-XLM-RoBERTa-Large	0.7803	0.4946
XLM-RoBERTa-Large	0.5964	0.4070
RoBERTalex	0.4641	0.3487
MEL	**0.8812**	**0.6485**

Table 3: Max F1 scores and learning speed for various models on the Multieurlex dataset. Bold values indicate the best scores.

6

5 Discussion and Conclusion

The results of this study highlight the importance of domain-specific pre-training for legal text classification, particularly in a language as resource-scarce as Spanish for legal corpora. By comparing our pretrained model, MEL, against both general-purpose multilingual models and existing domain-specific models, several key observations emerge.

- **Importance of Domain-Specific Pre-training**: The superior performance of MEL across both multilabel and multiclass classification tasks underscores the value of domain-specific pre-training. While general-purpose models like xlm-roberta-large can capture linguistic features across multiple languages, they lack the depth required for complex and technical vocabulary aspects present in legal texts. This observation aligns with findings in other domains, where pre-training on domain-specific corpora has proven to yield better results on task-specific evaluations.

- **Comparison to Other Domain-Specific Models**: While joelniklaus/legal-xlm-roberta-large is also a domain-specific model, its performance was consistently outpaced by MEL. This difference could be attributed to several aspects:

  - **Focus on Spanish Legal Texts**: Given that the purpose of our model was to achieve good results in Spanish, the focus on Spanish texts allowed the model to better understand language-specific features that the multilingual model may not capture with rigorous detail.

  - **Pre-training Corpus Quality**: By narrowing the corpus to Spanish language, a more meticulous cleaning was performed, providing a higher-quality dataset and allowing the model to learn from less corrupted data or noise.

- **Challenges in Spanish Legal Tasks**: Token or span classification tasks remain unevaluated due to the lack of annotated texts. This highlights the need for further development of labeled datasets in the legal domain.

5.1 Future Work

Future work could involve several directions to address the limitations and extend the applicability of this study:

- Developing annotated datasets for token or span classification tasks in Spanish legal texts to facilitate evaluation and fine-tuning.

- Expanding the pre-training corpus to include complex legal text extraction tasks, enabling models like MEL to handle more nuanced legal applications.

- Testing MEL on diverse downstream tasks such as question answering, information retrieval, and summarization within the legal domain.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] 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, 2020.

[4] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, 2019.

[5] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. Multilegalpile: A 689gb multilingual legal corpus, 2024.

[6] joelniklaus/legal-xlm-roberta-large · Hugging Face — huggingface.co. `https://huggingface.co/joelniklaus/legal-xlm-roberta-large`. [Accessed 22-11-2024].

[7] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[8] Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. A survey of spanish clinical language models, 2023.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

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

[12] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. Pre-trained language models for the legal domain: A case study on indian law, 2023.

[13] Rodrigo Agerri y Eneko Agirre. Lessons learned from the evaluation of spanish language models. *Procesamiento del Lenguaje Natural*, 70(0):157–170, 2023.

[14] Joel Niklaus. legal-xlm-roberta-large. `https://huggingface.co/joelniklaus/legal-xlm-roberta-large`, 2023. Multilingual legal language model pretrained on Multi Legal Pile.

[15] Joel Niklaus, Veton Matoshi, Matthias Sturmer, Ilias Chalkidis, and Daniel E. Ho. Multilegalpile: A 689gb multilingual legal corpus. *ArXiv*, abs/2306.02069, 2023.

[16] Jesus Cerquides Bueno and Paula Mateos Marín. Spanish congress parliamentary records (1977-2024), 2024.

[17] Council of the European Union European Parliament. Directive (eu) 2016/2102 of the european parliament and of the council of 26 october 2016 on the accessibility of the websites and mobile applications of public sector bodies (text with eea relevance). `http://data.europa.eu/eli/dir/2016/2102/oj`, 2016.

[18] Jeremy Singer-Vine and The pdfplumber contributors. pdfplumber, August 2024.

[19] Facebookresearch. Facebookresearch/fasttext: Library for fast text representation and classification.

[20] Trainer — huggingface.co. `https://huggingface.co/docs/transformers/main_classes/trainer`. [Accessed 20-12-2024].

[21] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer, 2021.

[22] PlanTL-GOB-ES/RoBERTalex · Hugging Face — huggingface.co. `https://huggingface.co/PlanTL-GOB-ES/RoBERTalex`. [Accessed 22-11-2024].