Computer Law & Security Review 52 (2024) 105908

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain

Daniele Licari *, Giovanni Comandè

Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà, 33, Pisa, 56127, Italy

ARTICLE INFO

Keywords:
Italian Legal NLP
Legal AI
Pre-trained Language Models
Italian Legal BERT

ABSTRACT

Legal-BERT models are based on the BERT architecture (or its variants) and have been developed specifically for the legal domain. They have reached the state of the art in complex legal tasks such as legal research, document synthesis, contract analysis, argument extraction, and legal prediction. In this paper, we proposed four versions of Legal-BERT models pre-trained on the Italian legal domain. They aim to improve NLP applications in the Italian legal context. We have shown that they outperforms the Italian "generalpurpose" BERT in several domain-specific tasks, such as named entity recognition, sentence classification, semantic similarity with Bi-encoders, and document classification.

1. Introduction

In many domains, specialized models performed better than pre-trained models on general domains [1–5]. In general, the more semantically distant a domain-specific language is from the common language than the greater the advantages of using specialized models, especially for complex tasks.

In the Italian legal context, the discrepancy between the specific language and the general language is even more pronounced. The Italian legal language has its unavoidable complexity, like all technical languages, but it is made even more obscure by stylistic expedients that often forcibly show a continuity with the languages of the past (Latin or old Italian). Full understanding of legal texts is the exclusive prerogative of experts in the field, who are able to grasp the meaning of these expressions. Carofiglio [6], a former Italian magistrate and politician, compared legal jargon to a foreign language that one learns in college to be admitted to a guild. A language capable of excluding the uninitiated from courtrooms and judicial acts. It can be perceived by software as a bizarre language that violates Italian grammar and syntax (characterized by imprecision, vagueness, opacity, stereotypes, archaisms, redundant circumlocutions, formular phrases, abuse of subordinates). It contains technical terms with specific and unambiguous meanings such as "contumacia", "anticresi", "anatocismo", and "sinallagma" (contumacy, antichresis, compound interest, reciprocity). It also makes extensive use of terms that are in common use, but are often used with their own and specific meanings, if not completely different from those in common use. For example, "nullita`", "annullabilita`", "inefficacia", "inutilizzabilita`" (nullity, voidability, ineffectiveness,inadmissibility), which outside of legal language are synonyms for annulment, refer to entirely different and distinct concepts and situations. Such locutions as "buon padre di famiglia" (good family man) and "possessore di buona fede" (possessor of good faith) indicate different concepts from the language of common use [7].

Traditional language models often struggle to overcome the limitations posed by the linguistic complexity of legal text. Their training does not adequately prepare them to understand and process the intricate nature of legal language, including its grammar, syntax, and specialized terminology. However, these limitations can be effectively addressed by employing models that are specifically trained and fine-tuned to handle legal text. By focusing on the unique characteristics and nuances of legal language, these specialized models offer improved accuracy and proficiency in comprehending and analyzing legal documents.

Chalkidis et al. [8] made significant contributions to the field of legal natural language processing by developing the LEGAL-BERT framework, which encompasses a family of transformer-based models tailored specifically for the English legal domain. Their research aimed to enhance the performance of the general-purpose BERT-BASE model in various prediction tasks by incorporating legal domain expertise. LEGAL-BERT consists of multiple variants, each distinct in terms of its pre-training methodology and the sources of training data. Chalkidis et al. employed two primary methods of domain adaptation within the LEGAL-BERT framework: 1. Additional Pre-training: In this approach, a

* Corresponding author.
E-mail address: d.licari@santannapisa.it (D. Licari).

https://doi.org/10.1016/j.clsr.2023.105908

Available online 28 October 2023

BERT-BASE model underwent further pre-training on a diverse range of legal text sources, such as EU legislation, UK legislation, US contracts, and US cases. This additional pre-training equipped the model with domain-specific knowledge, enabling it to understand and classify legal documents more effectively. 2. From-Scratch Pre-training: Another variant of LEGAL-BERT was pre-trained entirely from scratch using legal documents as the sole training data source. This model, starting with no prior knowledge from a general BERT model, learned directly from legal text corpora, including court cases, statutes, and legal articles. These variants of LEGALBERT were trained on diverse legal text sources, allowing them to capture the nuanced linguistic patterns, terminologies, and contextual cues unique to the legal domain. Chalkidis et al. conducted a series of experiments to evaluate the performance of LEGAL-BERT across various legal prediction tasks, such as legal document classification, legal entity recognition, sentiment analysis in legal texts, and legal question answering. The findings of their research demonstrated that LEGAL-BERT consistently outperformed the general-purpose BERT-BASE model across these tasks. The incorporation of domain-specific knowledge led to higher prediction accuracy, improved comprehension of legal language, and enhanced contextual representation for legal documents. The development of LEGAL-BERT bears substantial practical implications for the legal profession. It provides a powerful tool for legal research, document analysis, and information retrieval. Legal professionals can leverage LEGAL-BERT to efficiently categorize and analyze legal documents, identify relevant case law, extract legal concepts, and make more informed decisions.

Inspired by LEGAL-BERT, we present the development of ITALIANLEGAL-BERT models, capable of understanding the semantic meaning of Italian legal texts as never before.

In this paper, we make the following contributions:

1. We publicly release[1] on hunggingface hub [9] the ITALIAN-LEGAL BERT models to support Italian legal NLP research; to the best of our knowledge, they are the first language models pre-trained on a large corpus of Italian legal cases.
2. We proposed two versions of the ITALIAN-LEGAL-BERT model usingtwo different methods of domain adaptation: a model with additional pre-training of the Italian BERT-base model on Italian civil cases (ITALEGAL-BERT-FP) and its pre-trained variant from scratch on Italian legal documents (ITA-LEGAL-BERT-SC) based on the CamemBERT architecture.
3. We 'distilled' the knowledge of ITALIAN-LEGAL-BERT(-FP) on smallermodel (only 4 transformation layers instead of 12) three times faster than the teacher model. It is essential in all applications that work with a large amount of data (e.g., search engines and clustering on big data).
4. We created two variants of ITALIAN-LEGAL-BERT to handle long-sequences based on Local + Sparse + Global Attention. It provides flexibility in how the model attends to the input sequence. Local and sparse attention help address the computational complexity associated with attending to all positions, making it more feasible to process long sequences. On the other hand, global attention ensures that the model can capture global dependencies and consider the entire context when necessary.
5. We show that ITALIAN-LEGAL-BERT models outperform generalizedequivalents in several downstream tasks, such as sequence classification, document classification, semantic similarity, and named entity recognition in the Italian legal domain.

2. Related work

The legal writing system is very different from generic texts, with many domain-specific peculiarities. This discrepancy affects the pre-trained NPL models. Some researchers have demonstrated that the use of domain-specific pre-trained models can improve the performance of downstream tasks in the legal domain. Chalkidis et al. [8] proposed the LEGAL-BERT models pre-trained from scratch on 11.5 GB of legal texts and its variant further pre-training BERT-base model on legal corpora. Their experiments showed more substantial improvement in the most challenging end-task (i.e., multilabel classification in ECHR-CASES and contract header, lease details in CONTRACTS-NER), where in-domain knowledge is more important. Furthermore, no significant differences in performance were found between the two LEGAL-BERT variants.

Similar results are reported by Zheng et al. [10]. They also trained LEGAL-BERT models both with additional pre-training from the BERT base and with pre-training from scratch using a 37GB legal text collection. They compared their LEGAL-BERT and BERT-Base models on several downstream NLP tasks with different levels of difficulty and domain specificity. They suggest using domain-specific pre-trained models for very difficult legal tasks. They outperformed BERT-Base on complex downstream tasks such as identifying whether contract terms are potentially unfair [11]. In contrast, additional domain pre-training adds little value to simpler tasks compared to BERT-Base.

Recent works by Zhang et al. [12,13] on legal argument mining confirm this trend. Domain-specific BERT variants have achieved strong performance in many tasks. No significant differences were found between the two different methods of domain adaptation.

'Vanilla' Transformer models, such as LEGAL-BERT, may have difficulty capturing long-range relationships between words, and their use is limited to small portions of the text (typically 512 tokens) due to the computational cost of the attention mechanism. Condevaux and Harispe [14] proposed a new Local-Sparse-Global (LSG) attention mechanism to handle long sequences. Local attention allows the model to focus on specific parts of the input, sparse connections are used to extend the local context by selecting an additional set of tokens according to a set of rules (e.g. using norm, stride, pooling, LSH approach), while global attention allows the model to consider the relationships between different parts of the input. They demonstrated the superiority of the LEGAL-BERT with LSG attention mechanism over the original LEGAL-BERT model on three Multi-class classification legal tasks [15–17].

The success in this area encouraged researchers to build pre-trained language models on legal corpora in different languages [18]. Masala et al. 2021 [19] published the jurBERT model pretrained on a large Romanian legal corpus. It outperformed several strong baselines for legal judgment prediction tasks. In the same year, Douka et al. [20] created a language model adapted to French legal texts and showed that their model outperformed its generalized counterparts in the French legal domain. In China, researchers [21] have improved many predictive tasks on long Chinese legal documents using a pretrained language model on millions of documents published by the Chinese government. In 2022, AL-Qurishi et al. presented AraLegal-BERT [22], a pretrained language model for Arabic legal text, and showed that their model achieved better accuracy than the general BERT over the Arabic legal text. In Italy, A. Tagarelli and A. Simeri 2022 [23] proposed LamBERTa models for retrieving law articles, developing a BERT further pre-trained on the Italian Civil Code (ICC, few megabytes of data). Their model outperformed the ``predecessors" of BERT text classification models (BiLSTM, TextCNN, TextRCNN, Seq2Seq, Transformer) for prediction tasks on ICC articles.

In the same year, we introduced the ITALIAN-LEGAL-BERT model [24], which involved an extended pre-training process of the Italian BERT-base model using Italian civil cases. This current paper expands upon the initial ITALIAN-LEGAL-BERT paper presented at the conference and encompasses several noteworthy contributions. Primarily, we present the distillation of ITALIAN-LEGAL-BERT's knowledge into a smaller and more efficient model that operates at three times the speed while preserving its effectiveness. Additionally, we introduce a novel model based on CamemBERT, which undergoes comprehensive

[1] on huggingface.co/dlicari

pre-training from scratch on Italian legal documents. This new model greatly enhances the available resources for understanding legal language. To effectively handle long sequences, we introduce two variants of ITALIAN-LEGAL-BERT that incorporate a combination of local, sparse, and global attention mechanisms. These variants address the challenges associated with processing extended inputs and ensure improved performance. Furthermore, we enhance the evaluation of ITALIAN-LEGALBERT by expanding the dataset with additional data and models. This comprehensive evaluation allows for a thorough assessment of its capabilities across various downstream tasks. Additionally, we conduct comprehensive evaluations specifically focused on document classification tasks, further exemplifying the versatility and practicality of the proposed models. Overall, these contributions significantly advance the capabilities and practicality of ITALIAN-LE GAL-BERT, effectively establishing its effectiveness in processing legal language tasks.

3. ITALIAN-LEGAL-BERT models

In this section, we will delve into the implementation details of the ITALIANLEGAL-BERT models. To differentiate between the two domain adaptation approaches, we have labeled the model that underwent further pretraining on the Italian legal domain starting from the Italian BERT model as ITALIAN-LEGAL-BERT-FP, and the model pre-trained from scratch as ITALIAN-LEGAL-BERT-SC.

Background. BERT (Bidirectional Encoder Representations from Transformers) [25] is a contextual word embedding model using the transformers architecture [26] that can create context-sensitive embedding for each word in a given sentence, which will then be used for downstream tasks. BERT can be embedded in a downstream task and is developed as a task specific integrated architecture.

The Italian XXL BERT model, hereafter referred to as Italian BERT, is a cased model with 12 layers, 768 hidden units, 12 attention heads, and 110 million parameters. It follows the Bidirectional Encoder Representations from Transformers architecture and has been trained on a large Italian corpus of 81 GB. The training data includes Italian Wikipedia, texts from the OPUS corpora collection (opus.nlpl.eu), and data from the Italian portion of the OSCAR corpus (oscar-corpus.com). The Italian XXL BERT model is available on the Huggingface model hub[9] and was trained by the MDZ Digital Library team at the Bavarian State Library.

ITALIAN-LEGAL-BERT-FP. This model was previously introduced in our article [24] and it was created by continuing the training of Italian BERT on the target domain using masked language modeling. ITALIANLEGAL-BERT-FP is initialized with the Italian XXL BERT model and further pre-trained for an additional 4 epochs on 3.7 GB of Italian legal text sourced from the National Jurisprudential Archive.

The duration of the training phase was determined by carefully evaluating the specific characteristics of the dataset and the computational resources available. Firstly, the size of the dataset played a crucial role in the decision-making process. The dataset under consideration was of substantial magnitude, providing an ample amount of data to capture the necessary patterns and linguistic nuances inherent to the legal domain. The dataset size increased the likelihood of the model acquiring a comprehensive understanding of the complex language structures and legal terminology with a limited number of epochs. Additionally, the computational resources and time required for pre-training were taken into account. Training a largescale language model like BERT can be computationally intensive and timeconsuming. Given the constraints of the available resources, a careful balance needed to be struck between maximizing model performance and efficiently utilizing the resources at hand. Consequently, after careful consideration, it was determined that conducting the training for four epochs would be a practical and efficient decision. This choice ensured that the available resources were utilized effectively while still allowing the model to learn meaningful representations from the dataset. Overall, the duration of the training phase was determined by considering the dataset's size, the computational resources available, and the need to strike a balance between efficiency and model performance. This approach enabled the achievement of a robust training process that effectively captured the intricacies of the legal domain within the given resource constraints.

During the training process, we utilized the BERT architecture with a language modeling head on top and employed a dynamic masking pattern[27] with WordPiece tokenization[28]. The adoption of a dynamic masking pattern, particularly in conjunction with WordPiece, can provide several advantages. Unlike a fixed masking pattern, dynamic masking randomly masks different tokens in each training instance. This approach encourages the model to learn more robust and general representations of the language, as it must predict the masked tokens based on the surrounding context.

To optimize the training, we used the AdamW optimizer and set an initial learning rate of 5e-5 with linear decay. The sequence length was set to 512, ensuring that the model could effectively capture long-range dependencies within the text. Due to GPU capacity constraints, we chose a batch size of 10 for training. Over the course of training, we performed 8.4 million training steps. The training was conducted on a single GPU V100 with 16GB of memory. For more detailed information regarding the specific hyperparameters employed in the training process, please refer to the appendix in Table A.10.

ITALIAN-LEGAL-BERT-SC. The ITALIAN-LEGAL-BERT-SC model was trained from scratch using a larger training dataset consisting of 6.6GB of civil and criminal cases. To achieve this, we employed the CamemBERT architecture [29] with a language modeling head on top. During training, we utilized the AdamW optimizer with an initial learning rate of 2e-5 and linear learning rate decay. The sequence length was set to 512, and a batch size of 18 was used. To handle the computational demands of the training process, we leveraged the power of 8 NVIDIA A100 GPUs, each equipped with 40GB memory. By employing distributed data parallelism, we were able to process 8 batches per training step, significantly enhancing the training throughput. For more detailed information regarding the specific hyperparameters used during training, please refer to the appendix in Table A.11.

While ITALIAN-LEGAL-BERT-SC shares similarities with the previous model, there are notable differences. Instead of WordPiece tokenization, we opted for SentencePiece tokenization [30], which was trained from scratch on a subset of the training set, consisting of 5 million sentences. The vocabulary size was set to 32,000, ensuring a comprehensive coverage of the legal language. The sentencepiece tokenizer is a widely adopted subword tokenization method known for effectively handling out-of-vocabulary words, capturing morphological variations, and reducing the overall vocabulary size. In the tokenization process, the text is divided into a sequence of subword units, taking into account both the boundaries of individual words and the subword units within words. The main distinction between WordPiece and SentencePiece lies in their vocabulary generation. WordPiece constructs its vocabulary by considering the frequency of subwords in the training corpus. In contrast, SentencePiece employs unsupervised training algorithms, such as byte-pair encoding (BPE), for vocabulary generation. This algorithm analyzes the text's statistical properties and identify frequently occurring subword units. Through the training process, the algorithm learns patterns in the text and determines the optimal subword units to form a vocabulary. This vocabulary comprises the identified subword units, which are then treated as tokens during both model training and inference stages. By using subword units as tokens, the model can capture fine-grained linguistic information and effectively represent complex structures within the text. This integration enables the ITALIAN-LEGAL BERT model to achieve a more comprehensive and accurate representation of the Italian legal language An advantageous aspect of utilizing the CamemBERT implementation from the Huggingface transformers library is its seamless integration of the sentencepiece tokenizer with the ITALIAN-LEGAL BERT model.

Training Datasets. To train ITALIAN-LEGAL-BERT-FP, we initially used a dataset consisting of 235,000 documents specifically focused on civil law cases. Subsequently, we expanded the dataset by incorporating approximately 300,000 documents related to criminal law cases to train ITALIANLEGAL-BERT-SC from scratch, encompassing both civil and criminal contexts. It is important to note that the legal documents used in this study were obtained through scientific collaboration agreements with various Italian courts, making them unavailable for public distribution.

The decision to adapt ITALIAN-LEGAL-BERT-FP solely to the civil context was driven by the overarching objectives of the primary project, predictivejustice.eu, which aims to enhance the efficiency of civil justice. As a result, the focus of this specific model was tailored to civil law cases. Conversely, for the ITALIAN-LEGAL-BERT-SC model, we chose to include all available data, including criminal law cases. Training the model from scratch poses a higher risk of overfitting, and incorporating a diverse range of examples helps the model generalize better across different legal contexts.

Preprocessing Datasets. To prepare the legal documents for training, we employed various preprocessing steps. First, we utilized the Tika framework [31] to convert the documents from their original format to plain text. Subsequently, we applied several regular expression-based cleaning functions to the law corpora, aiming to consolidate whitespace and remove unnecessary line breaks. For sentence segmentation, we customized the spaCy model for the Italian language by incorporating new tokenization rules. These rules specifically targeted abbreviations and acronyms commonly found in Italian legal texts.[2] By incorporating these exceptions, we enhanced the accuracy of sentence segmentation in legal documents. The segmented sentences underwent further cleaning, where we removed special characters and filtered out sentences that were too short (less than five words), ensuring the quality and relevance of the training data. We implemented a filtering process to remove sentences with less than five words. This decision was based on the understanding that very short sentences lack sufficient context and may not contribute meaningful information for training a language model. By excluding such sentences, we aimed to enhance the overall quality and coherence of the training data.

The performance of language models is often improved when trained on sentences of a certain minimum length. Longer sentences provide more contextual information and exhibit a greater syntactic structure, enabling the model to capture more intricate linguistic patterns and dependencies. On the other hand, very short sentences, particularly those consisting of just a few words, are more likely to be noisy or incomplete fragments of text. Consequently, by filtering out these sentences, we effectively reduced noise within the training data, thereby enhancing the overall signal-to-noise ratio.

The preprocessed dataset used for training the ITALIAN-LEGAL-BERTFP model consisted of 21,004,500 sentences and 498,002,402 words, amounting to 3.7 GB of data. In contrast, the final corpus employed for ITALIANLEGAL-BERT-SC encompassed 33,276,194 sentences and 855,452,531 words, totaling 6.6 GB.

Finally, to prepare the input data for the ITALIAN-LEGAL-BERT-FP model, we applied the Italian BERT tokenizer to tokenize the sentences from the corpus. The sentences were truncated to a maximum length of 512 tokens. Conversely, for ITALIAN-LEGAL-BERT-SC, we utilized our pretrained SentencePiece tokenization specifically trained on a subset of Italian legal documents.

MLM Evaluation. Perplexity (PPL) is a commonly used metric for evaluating language models. In the context of the Masked Language Modeling (MLM) objective, perplexity is computed by making predictions for the masked tokens in separate evaluation datasets, while having access to the remaining tokens. In our evaluation, we utilized two distinct datasets: one consisting of 20,000 civil cases and another containing 20,000 criminal cases. These datasets were excluded from the training process and underwent the same pre-processing procedure as the training set. The evaluation corpus for civil cases comprised 592,426 sentences and 35,378,854 words, while the corpus for criminal cases comprised 506,831 sentences and 18,941,252 words. To calculate perplexity, a portion of the tokens in the evaluation datasets were replaced with special mask tokens. Specifically, these mask tokens accounted for 15% of the total tokens in the evaluation datasets. The language model was then tasked with predicting the original tokens that were masked based on the surrounding context. By comparing the predicted tokens with the original tokens, we computed the cross-entropy loss for the masked tokens. This loss quantifies how well the language model can estimate the probability distribution of the masked tokens given the context. Formally, the perplexity (denoted as PPL) of an MLM model on a set of masked tokens in a text corpus is defined as:

$$PPL(D) = exp\left(-\frac{1}{N} \sum_{n}^{n=1} \log(P(w_i|context)) \right)$$

In this formula:

- D represents the dataset containing examples with masked tokens that the model is evaluated on.
- N is the total number of masked tokens in the dataset.
- $P(w_i|context))$ is the probability assigned by the MLM model to the correct word w_i given the context (the surrounding words) in which w_i appears.

It's important to emphasize that the evaluation datasets used for computing perplexity were separate from the training data. These datasets serve as a measure of the language model's generalization ability to unseen examples. By evaluating the model's accuracy in predicting the masked tokens, we gain insights into its effectiveness in capturing underlying language patterns and generating coherent and contextually appropriate text.

The findings presented in Table 1 highlight some interesting observations regarding the performance of the ITALIAN-LEGAL-BERT models compared to the general BERT model. Specifically, the ITALIAN-LEGAL-BERT-FP model demonstrates a notable reduction in evaluation perplexity when applied to civil cases, indicating an improvement in its language modeling capabilities within the civil law domain. However, in the context of criminal cases, the general BERT model outperforms both the ITALIAN-LEGAL-BERTFP and ITALIAN-LEGAL-BERT-SC models. It is worth noting that the ITALIAN-LEGAL-BERT-SC model shows a slight improvement in perplexity compared to the ITALIAN-LEGAL-BERT-FP model for criminal cases, albeit not reaching the performance level of the general BERT model.

To delve deeper into the analysis, a qualitative investigation was conducted by seeking input from civil law judges. They were presented with domain-specific sentences, and the models were tasked with predicting the masked words within the legal context. Unlike the previous random masking approach, the masking was done strategically based on the legal context to focus on significant words. The results presented in Table 2 demonstrate the models' ability to infer the masked words, with

Table 1
Perplexity scores on evaluation datasets.

Dataset	#Sentences	Model	PPT
Civil cases	592,426	ITALIAN-LEGAL-BERT-FP	**3.4580**
		ITALIAN-LEGAL-BERT-SC	6.1472
		Italian BERT	4.5430
Criminal cases	506,831	ITALIAN-LEGAL-BERT-FP	4.9520
		ITALIAN-LEGAL-BERT-SC	4.7733
		Italian BERT	**4.3137**

[2] The complete list can be accessed at https://huggingface.co/dlicari/Italian-LegalBERT/blob/main/abbreviazioni.csv

Table 2
Results of the Italian BERT and ITALIAN-LEGAL-BERT mask filling pipeline on the prediction of a single mask (strikethrough words). The probability of a specific token is reported in parentheses.

Sentence (Mask is strikethrough)	ITA BERT	ITA-LEGAL-BERTFP	ITA-LEGAL-BERT-SC
Il **padre** pu`o vedere il figlio a weekend alternati en: The **faather** can see his son on alternate weekends	genitore (53.61%) **padre** (27.70%) pap`a (6.81%) marito (2.19%) proprietario (0.62%)	**padre** (99.24%) genitore (0.56%) ricorrente (0.05%) resistente (0.03%) pap`a (0.03%)	**padre** (97.37%) genitore (1.71%) ricorrente (0.35%) coniuge (0.10%) resistente (0.07%)
viene **stabilita** una collocazione paritetica dei figli en: an equal placement of the children is **established**.	garantita (24.1%) meno (10.72%) proposta (6.66%) **stabilita** (4.52%) assicurata (4.09%)	prevista (40.48%) **stabilita** (21.81%) disposta (12.74%) assicurata (6.32%) garantita (1.77%)	**stabilita (31.69%)** prevista (14.73%) disposta (13.80%) fissata (4.35%) prospettata (2.26%)
assegno di mantenimento comprensivo di spese **straordinarie** en: maintenance allowance including **extraordinary** expenses.	. (38.58%) mediche (17.01%); (6.62%) legali(4.55%) generali (3.35%)	**straordinarie** (69.25%) : (7.61%) extra (4.86%) mediche (4.30%) . (4.20%)	mediche (24.51%) . (20.28%) extra (14.73%) **straordinarie** (12.48% : (8.13%)
viene stabilito il **mantenimento** diretto en: direct **maintenance** is established	trattamento (8.58%) prezzo (7.43%) contratto (5.08%) contributo (4.23%) lavoro (4.06%)	pagamento (48.93%) versamento (23.89%) **mantenimento** (5.20%) trasferimento (2.54%) rimborso (2.09%)	pagamento (11.67%) compenso (9.41%) rapporto (5.71%) **mantenimento** (3.52%) collegamento (3.19%)
cambiamento di **sesso** senza operazione chirurgica en:**sex** change without surgery	pelle (19.12%) capelli (16.54%) **sesso** (8.53%) colore (6.17%) peso (4.48%)	**sesso** (89.01%) genere (5.40%) nome (1.20%) profilo (1.01%) persona (0.43%)	destinazione (10.51%) **sesso** (8.89%) diagnosi (7.22%) genere (3.93%) denominazione (3.81%)
Il **ricorrente** ha chiesto revocarsi l'obbligo di pagamento. en: The **plaintiff** requested that the payment obligation be revoked.	Comune (11.89%) giudice (9.17%) cittadino (4.70%) lavoratore (3.17%) sindaco (2.62%)	**ricorrente** (72.64%) convenuto (9.64%) resistente (3.99%) lavoratore (2.90%) Ministero (2.53%)	**ricorrente** (65.29%) convenuto (3.80%) richiedente (3.60%) Condominio (2.39%) lavoratore (2.09%)
Non avendo la **Corte** di merito valutato la prova en: Not having the **Court** of merit assessed the eviden	Comune (11.89%) giudice (9.17%) cittadino (4.70%) lavoratore (3.17%) sindaco (2.62%)	**Corte** (56.29%) corte (23.26%) sentenza (12.49%) giurisprudenza (3.87%) decisione (1.64%)	**Corte** (88.25%) corte (10.91%) sentenza (0.66%) decisione (0.06%) pronuncia (0.02%

the correct word consistently appearing among the top suggestions generated by both the ITALIAN-LEGAL-BERT-FP and ITALIAN-LEGAL-BERT-SC models.

However, it is important to recognize that improvements in intrinsic perplexity do not necessarily guarantee improvements in extrinsic performance for specific language processing tasks. The ability to accurately recognize the context of entire sentences or identify domain-specific terms may be more critical in certain tasks than the recognition of high-frequency tokens considered in perplexity computation using random masking, such as stopwords. To gain a comprehensive understanding of the performance of the ITALIAN-LEGALBERT models, we will conduct a series of experiments on downstream tasks in the forthcoming ``Downstream Evaluation Task" section. These experiments will assess the models' performance in specific language processing applications, providing a more comprehensive analysis of their relative and absolute performance behaviors compared to the general BERT model.

By evaluating the ITALIAN-LEGAL-BERT models' performance in downstream tasks, we aim to determine their practical utility in real-world scenarios and shed light on their strengths and weaknesses. This comprehensive evaluation will enable us to gain a deeper understanding of how well these models capture the underlying language patterns and generate coherent and contextually appropriate text within the legal domain.

3.1. DISTIL-ITALIAN-LEGAL-BERT

"Knowledge distillation" is a technique in which a large, complex model (known as the ``teacher model") is used to train a smaller, more streamlined model (known as the ``student model"). The goal of knowledge distillation is to transfer the knowledge and expertise of the teacher model to the student model, allowing the student model to achieve similar performance while being more efficient in terms of computational resources and memory requirements.

The motivation behind knowledge distillation lies in the desire to create models that are more lightweight and suitable for deployment on resourceconstrained devices or in situations where computational efficiency is crucial. By distilling the knowledge from a larger teacher model into a smaller student model, we can obtain a model that is more efficient, faster in inference, and requires fewer computational resources. This makes it feasible to deploy the model on devices with limited memory and processing power, opening up possibilities for real-time applications and scenarios where quick responses are essential.

In the context of the Italian legal language, knowledge distillation can bring several key advantages. Firstly, the legal domain often deals with large volumes of complex text, making the use of computationally demanding models challenging in practice. By employing knowledge distillation, we can develop student models that are specifically tailored to the Italian legal language and its unique characteristics, while still benefiting from the expertise of larger and more comprehensive teacher models.

So, We used the process of knowledge distillation to create a fast, lightweight student model with only four levels of Transformers, capable of producing sentence embeddings similar to those produced by the more complex ITALIAN-LEGAL-BERT-FP teacher model. The student model, DISTILITALIAN-LEGAL-BERT, was initialized with four layers of Transformer taken from the teacher model and optimized using Sentence-BERT [32] library by minimizing the mean square error (MSE) between its embeddings and those produced by the teacher model.

It was trained on the ITALIAN-LEGAL-BERT-FP train set (3.7 GB) for 4 epochs using the AdamW optimizer, initial learning rate 1e-4 (with linear decay), batch size 24, warm-up steps 1000, evaluation step 5000, and automatic mixed precision (AMP) (details in the appendix tab A.12). Once the distillation process was completed, the student model was able to produce embeddings similar to those produced by the teacher model but much faster. In our experiments presented in the 'Downstream evaluation task 'section, we found that the DISTIL-ITALIAN-LEGAL-BERT with 4 Transformer layers was able to maintain a similar performance to the teacher model while being 3 times faster.

3.2. LSG-ITALIAN-LEGAL-BERT

The inclusion of the Local-Sparse-Global (LSG) attention mechanism in the ITALIAN-LEGAL-BERT models addresses the challenge of

effectively processing long texts. By integrating local and sparse attention, these models are capable of efficiently handling extensive legal documents while maintaining reasonable computational efficiency. This enhancement extends the applicability of ITALIAN-LEGAL-BERT to tasks that involve lengthy legal texts, such as document classification, summarization, and information extraction.

The key advantage of the LSG attention mechanism lies in its ability to capture both global dependencies and local context within long sequences. By incorporating global tokens and leveraging sparsity factors, the models can attend to relevant information throughout the entire document while simultaneously processing local neighborhoods efficiently. This combination of global and local attention empowers the models to capture crucial legal language patterns, dependencies, and semantic relationships across different sections of the text, thus improving their understanding and performance on legal text-related tasks.

Additionally, the LSG attention mechanism provides an effective solution for reducing computational complexity when dealing with long sequences. By employing sparsity patterns and selecting tokens based on norms, attention computations can focus on the most informative segments of the sequence, reducing overall computational requirements without sacrificing performance. This feature is particularly advantageous in resource-constrained environments or scenarios that necessitate real-time processing of lengthy legal documents.

To implement the LSG attention mechanism, we utilized the LSG converter script, available at https://github.com/ccdv-ai/convert checkpoint to lsg, which replaced the full attention mechanism in the encoder section of the ITALIAN-LEGAL-BERT-FP and ITALIAN-LEGAL-BERT-SC models. The

LSG attention configuration involved a maximum sequence length of 16,384, 7 global tokens, 128 local block size, 128 sparse block size, 2 sparsity factors, and a 'norm' sparse selection pattern that selected the tokens with the highest norms. These settings were carefully selected to strike a balance between capturing long-range dependencies, preserving local context, and achieving computational efficiency.

To evaluate the performance of the LSG variants, we conducted experiments specifically on the legal document classification task. The results demonstrated exceptional performance compared to the versions with the full attention mechanism constrained to 512 tokens. For more detailed insights and findings from these experiments, please refer to the Legal Document Classification section of this study.

In conclusion, the integration of the LSG attention mechanism into the ITALIAN-LEGAL-BERT models offers significant advantages for effectively handling long legal texts. The LSG variants exhibit excellent performance in legal document classification, surpassing the versions with full attention restricted to 512 tokens. By incorporating local and sparse attention, these models can efficiently process extensive legal documents, capture global dependencies, and preserve local context. Moreover, the LSG attention mechanism enhances computational efficiency, making it a valuable enhancement for deploying ITALIAN-LEGAL-BERT models in real-world applications that involve lengthy legal texts.

4. Downstream evaluation tasks

The Italian BERT and ITALIAN-LEGAL-BERT models were evaluated and compared on four domain-specific downstream tasks. In the first task, we trained the models with an additional sequence tagging layer on the top using spaCy[33] to recognize name/role of actors involved in the trial. For the second task, we fine-tuned the models with a sequence classification head for the sentence type classification. In the third downstream task, we tested the models on textual semantic similarity using sentence embeddings (mean pooling on the last layer of the models) and cosine similarity. In the last task, we tested them on the task of classifying legal documents based on their content and also evaluated LSG variants of the ITALIAN-LEGAL-BERT models.

4.1. Named entity recognition

We trained and evaluated the Italian-BERT and ITALIAN-LEGAL-BERT models on a Named Entity Recognition (NER) task to identify named entities based on the type of person found in the judgments. We defined 7 entity types, as shown in Table 3.

Dataset. We selected 118 judgments from the civil law database of the Court of Genoa, with which we have a scientific collaboration agreement. Given the considerable experience of our research group in these matters, the selected judgments were all personal injury judgments (no. 59) contained in the database, and an equal number of family judgments were selected stratified by text length. We then converted the PDF files to plain text using Tika [31], applied some text cleaning functions (removal of multiple blank lines and extra spaces), and converted the texts to an annotable data structure (jsonl format) for import into the Doccano annotation tool [34]. We set up and used the Doccano tool for quick and easy manual annotation of texts with the 7 predefined entities. The experts found and annotated 6355 entities; Table 3 shows the distribution of entities in the dataset. Finally, the dataset was divided into 80% for model training (10% of the training set for validation) and 20% for model evaluation in a stratified manner to preserve the distribution of entities in the two subsets.

Model architecture. We created our NER models using spaCy's v3.2 Named Entity Recognition system [33]. The model architecture in Fig. 1 consists of a two-tier pipeline: the contextual embedding layer and the transition-based chunking model [35]. The first tier of our NER models is the contextual embedding layer, responsible for encoding tokens into continuous vectors based on their context. This layer utilizes pre-trained language models, such as Italian BERT or ITALIAN-LEGAL-BERT, to capture the semantic meaning and representation of tokens within their surrounding context. The contextual embeddings provide rich contextual information that helps in understanding the nuanced characteristics of named entities. The second tier of our NER models is the transition-based chunking model. This model predicts the structure of the text by mapping it onto a set of state transitions. It uses the contextual word embeddings obtained from the previous layer to incrementally construct states from the input sequence. These states represent different parts of the text and are assigned entity labels using a multilayer neural network.

Training procedure. We trained four named entity recognition pipelines,

Italian BERT + spaCy's NER and ITALIAN-LEGAL-BERT variants+ spaCy's NER, using AdamW Optimizer, initial learning rate 5e-5 (with linearly decay), 20,000 maximum number of steps, 250 warm-up steps, early stopping patience on the F1 validation score, and batch size 128 (see Table A.13 in the Appendix for more details).

Evaluation. We compared the four NER pipelines using the exact match criterion with gold-standard entities (both entity boundary and type are correct) in the test set. F-score is used to evaluate and compare the performance. The results in Table 4 show that the NER pipeline with

Table 3
Entity type description and their distribution on train and test set.

Entity	Description	#Train	#Test
Person	Names of the main actors (plaintiffs and defendant)	3103	771
Person-Judge	Names of the judges	328	80
Person-Lawyer	Names of the lawyers	334	74
Person-Witness	Names of the witness	201	57
Person-Expert	Names of the experts (e.g., doctors, engineers)	136	24
Person-Family	Name of the family members of the plaintiffs and defendants	637	162
Person-Family-Children	Names of the children of the plaintiffs and defendants	353	95

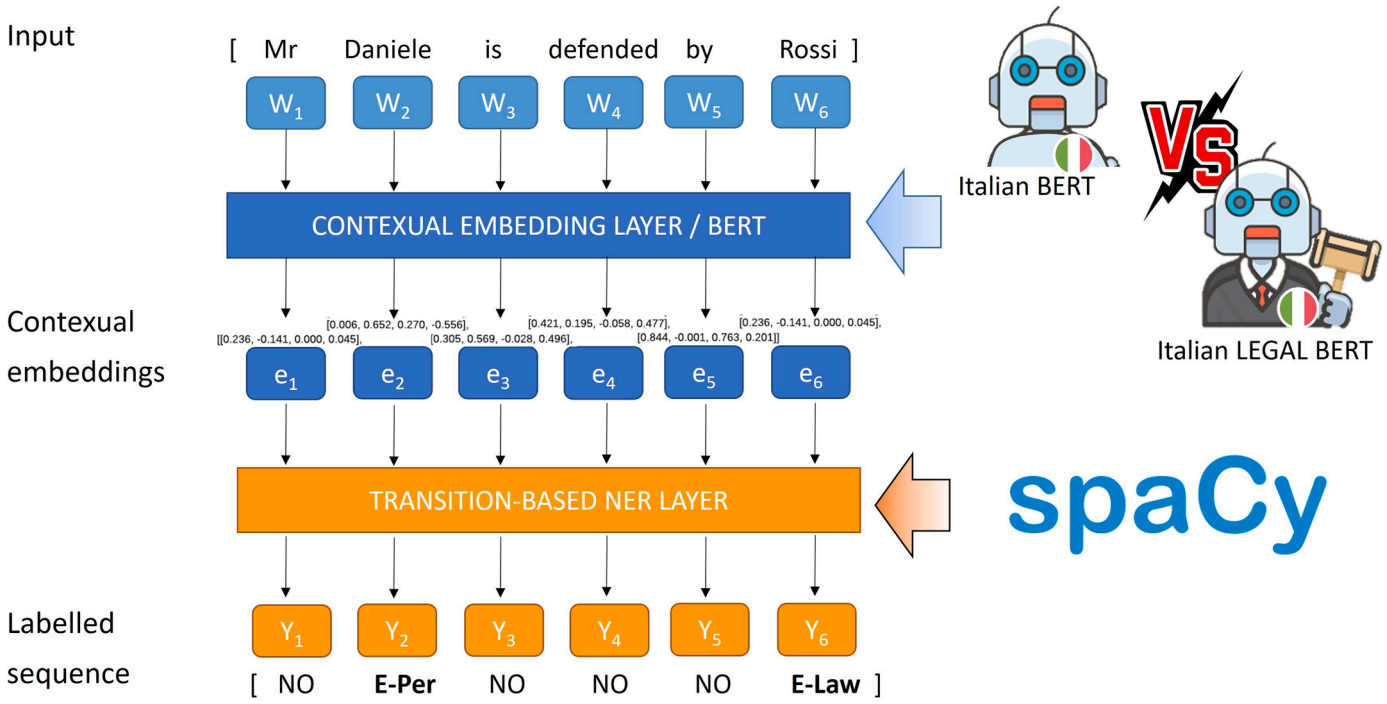

Fig. 1. spaCy-transformer's neural architecture for NER.

ITALIAN-LEGAL-BERT-FP contextual encoder and its distilled version outperforms that with Italian BERT in recognizing most entities. In our experiment, we found that a distilled model is able to maintain approximately 98.9% of the performance of the teacher model while being 2.7 times faster.

Notably, ITALIAN-LEGAL-BERT-SC is trained from scratch on the Italian legal domain. While this allows the model to be highly focused on the legal domain, it may not benefit as much from the general knowledge acquired by models pre-trained on a broader range of text. As a result, ITALIANLEGAL-BERT-SC may not perform as well on tasks that require a broader understanding of language or rely on general knowledge. The distilled model, DISTIL-ITA-LEGAL-BERT, offers a good compromise between performance and speed of inference and training. It is a compact version of the original model and is more efficient in terms of memory usage and computation time. The distilled model can be particularly useful in scenarios with limited computational resources or when processing large amounts of data. It maintains a reasonable level of performance while providing faster training and inference times compared to the full-sized models. In cases where the task is relatively simple and closer to general knowledge rather than highly complex legal-specific knowledge, using the distilled model or even a ITALIANLEGAL-BERT-FP may be more suitable.

This is because the task may not require the highly specialized domain knowledge and fine-grained understanding of legal language that the ITALIAN-LEGAL-BERT-SC models offer. Therefore, the choice of the best model depends on the complexity of the task, the availability of domain-specific knowledge, and the trade-off between performance and computational efficiency.

4.2. *Sentence classification*

Unlike the English legal context, there are no public datasets on which to test models on downstream NLP tasks in the legal context. Then, we created a new benchmark dataset for sentence classification tasks. A common civil judgment has 5 basic parts:

1. INTRODUCTION: an indication of the judge who pronounced it; anindication of the parties and their lawyers;
2. CONCLUSION OF THE PARTIES: the conclusions of the prosecutor (if any) and those of the parties;
3. DEVELOPMENT OF THE TRIAL: summary of the appealed judgment and reasons of appeal;
4. REASON: the concise statement of the factual and legal reasons forthe decision (the statement of reasons);

Table 4
F-score (F1) for The ITALIAN LEGAL BERT variants+Spacy NER and ITALIAN BERT+Spacy NER models evaluated using exact match criterion on individual entities, f1 macro-average, and pipelines speed in processed words per second on the test set (higher is better).

	ITA-BERT	ITA-LEGAL-BERT-FP	ITA-LEGAL-BERT-SC	DISTIL-ITA-LEGAL-BERT
Person-Judge	98.14	**98.77**	96.30	98.14
Person	82.45	**83.30**	79.29	82.01
Person-Lawyer	88.00	**91.50**	91.28	83.66
Person-Family-Children	50.63	**59.38**	38.55	47.40
Person-Witness	65.91	66.00	**80.00**	70.48
Person-Expert	69.23	67.69	62.75	**70.00**
Person-Family	10.75	8.47	8.08	**18.35**
AVG	66.45	**67.87**	65.18	67.15
SPEED	1911	1856	2652	**4981**

5. CONCLUSION: the decisional content of the judgment.

We want to evaluate the ITALIAN-LEGAL-BERT models on a sentence classification task by trying to predict the belonging section. Although this downstream task was created to benchmark it could have practical utility because Italian judgments do not follow a precise standard, often sections are merged or are identified in a variety of headers that making it difficult to apply rules based on regular expressions.

Dataset. We extracted 152,368 sentences from documents with 5 sections (using regular expression) from 1503 Italian Civil cases stratified on section length, Table 5.

Finally, the dataset was split 80% for training models and 20% for model evaluation (10% for validation and 10% for testing) in a stratified fashion on the section name to preserve the distribution of sentences across both subsets. The data were obtained through scientific collaboration agreements between some Italian courts and the Scuola Superiore Sant'Anna.

Classification Models. We trained Italian BERT and ITALIAN-LEGALBERT models with different sequence classification heads on top of the BERT output:

- BERT+Linear, it adds a Dropout layer on BERT pooled output with a Dropout rate 0.1 and an output Linear layer (BERT+Linear). Finally, the cross-entropy loss weighted class distribution of the training set is computed between the output of the last layer and target labels. The class weights are estimated from the class weight function of Sklearn [36] for unbalanced datasets.
- BERT+MLP, it adds two fully connected layers (FC) followed by REctify Linear Unit (RELU), a Normalization layer, an output Linear layer, and weighted cross-entropy loss.
- BERT+CNN, it adds three 1D convolution layers (Conv1d) incremental kernel sizes (3,4,5). Each layer is followed by RELU. A maximum global pooling layer is applied to the last convolution layer, followed by a Dropout layer (rate 0.1), an output Linear layer, and weighted cross-entropy loss.
- BERT+LSTM, it adds a bidirectional Long short-term memory (LSTM) layer on BERT sequence output, a Dropout layer with rate 0.1, an output Linear layer, and weighted cross-entropy loss.

Training procedure. We trained Italian BERT and ITALIAN-LEGALBERT models with different sequence classification heads on top (Linear, MLP, LSTM, and CNN layer) using the same hyperparameters configuration (Table A.14 in the Appendix) for a fair comparison between models and architectures.

Evaluation. The performance of the models was evaluated by Matthews Correlation Coefficient (MCC) scores in the test using the combination of four encoders and four classifiers. MCC is a metric commonly used to evaluate the performance of classification models in multi-class classification tasks. It extends the use of MCC from binary classification to handle multiple classes. In multi-class classification, MCC measures the quality of the overall classification, taking into account true positives, true negatives, false positives, and false negatives for each class. It provides a balanced measure that considers the performance across all classes and provides an aggregate score that reflects the model's ability to classify instances correctly across multiple categories.

Table 5
Distribution of sentences over the 5 sections.

SECTION NAME	N. SENTENCES
INTRODUCTION	11,737
CONCLUSION OF THE PARTIES	18,684
DEVELOPMENT OF THE TRIAL	32,174
REASON	78,701
CONCLUSION	11,072
TOTAL	152,368

Table 6
MCC and speed for sentence classification task on the test set using Italian BERT and ITALIAN-LEGAL-BERT variants with different classifiers on top (Linear, MLP, LSTM, CNN).

Encoder	Classifier	MCC	Train samples/s
ITA-BERT	Linear	0.809	93.67
	MLP	0.806	93.52
	LSTM	0.804	78.88
	CNN	0.809	88.02
ITA-LEGAL-BERT-FP	Linear	0.813	93.74
	MLP	0.813	93.64
	LSTM	0.812	79.08
	CNN	0.814	88.11
ITA-LEGAL-BERT-SC	Linear	0.812	95.77
	MLP	**0.815**	95.75
	LSTM	0.808	80.75
	CNN	0.813	89.91
DISTIL-ITA-LEGAL-BERT	Linear	0.809	261.28
	MLP	0.811	**261.34**
	LSTM	0.808	178.60
	CNN	0.811	221.52

The Table 6 presents the results of the MCC (Matthews Correlation Coefficient) and speed for a sentence classification task on the test set using different classifiers on top of Italian BERT and ITALIAN-LEGAL-BERT variants (FP and SC). The MCC score reflects the performance of the models, with higher values indicating better performance.

For Italian BERT, all the classifiers show comparable performance, with MCC scores ranging from 0.804 to 0.809. The Linear and CNN classifiers achieve the highest MCC scores of 0.809, while the MLP and LSTM classifiers achieve slightly lower scores of 0.806 and 0.804, respectively. In terms of speed, the LSTM classifier exhibits the slowest processing time, with 78.88 train samples per second, while the Linear and CNN classifiers demonstrate relatively faster speeds of 93.67 and 88.02 train samples per second, respectively.

When considering ITALIAN-LEGAL-BERT-FP, we observe the MCC scores range is from 0.812 to 0.814, with the MLP classifier achieving the highest score of 0.813. The Linear classifier also performs well, with a score of 0.813. In terms of speed, the ITALIAN-LEGAL-BERT-FP models demonstrate similar performance to Italian BERT.

Moving on to ITALIAN-LEGAL-BERT-SC, we observe a slight improvement in performance compared to the other variants. The MCC scores range from 0.808 to 0.815, with the MLP classifier achieving the highest score of 0.815. The Linear and CNN classifiers also demonstrate strong performance, with scores of 0.812 and 0.813, respectively. In terms of speed, ITALIANLEGAL-BERT-SC models exhibit similar processing times compared to the other previous variants, with the MLP and Linear classifiers achieving the highest speeds of 95.75 and 95.77 train samples per second, respectively.

Finally, the DISTIL-ITA-LEGAL-BERT variant shows comparable performance to ITALIAN-LEGAL-BERT-FP, with MCC scores ranging from

0.808 to 0.811. The MLP and CNN classifiers achieve the highest scores of 0.811. In terms of speed, the DISTIL-ITA-LEGAL-BERT models exhibit faster processing times compared to the other variants, with the MLP classifier being the fastest at 261.34 train samples per second.

In summary, the results indicate that ITALIAN-LEGAL-BERT variants, particularly ITALIAN-LEGAL-BERT-SC, offer improved performance compared to Italian BERT in terms of MCC scores. The choice of classifier on top of the models also influences performance, with the MLP classifier generally has a good trade-off between higher MCC scores and speed among the different variants. Noteworthy are the results of the distilled model, which outperforms the Italian Bert and is 3 times faster.

4.3. Semantic similarity with bi-encoders

We tested the ability of the model on the task of determining whether two pieces of legal text are similar, in terms of meaning. The strong assumption is that two contiguous sentences within a specific section are semantically related and refer to the same context, instead, two sentences taken randomly from two different documents and different sections can refer to a different context.

Dataset. We built the dataset by taking, from 1503 civil law judgments, pairs of contiguous portions of the text (of 5 sentences) in the "CONCLUSION OF PARTIES" and "DEVELOPMENT OF PROCESS" sections and text pairs from two different documents and sections. We labeled as 'similar' the contiguous pairs from the same document and 'unsimilar' the pairs from different documents. The final dataset contains 4018 text pairs (2009 labeled as 'similar' and the other 2009 as 'unsimilar'). The choice of taking similar sentences from the two sections was made on the basis that the "CONCLUSION OF THE PARTIES" and "DEVELOPMENT OF THE PROCESS" sections are more descriptive and with self-contained concepts than other sections such as "REASON" or "CONCLUSION" that contain many references to the previous sections. The initial assumption on the similarity has been validated by experts on a subset of the dataset.

Similarity Procedure. The semantic similarity between the text pairs in the dataset was evaluated using the Italian BERT and ITALIAN-LEGALBERT models to obtain the context vectors of the two sentences to be compared (using mean pooling on the last layer) and, then, the cosine similarity for the similarity scores between the pair of vectors ($\cos(x,y) = \frac{xy}{\|x\|\cdot\|y\|}$). For each model, a similarity threshold was established to identify similar and non-similar texts. The Fig. 2 shows the similarity scores distribution over the groups of 'similar' and 'unsimilar' pairs of sentences, calculated using the Italian BERT and ITALIAN-LEGAL-BERT models as contextual sentence encoders.

Optimized threshold. A similarity threshold is a numerical value that is applied to the similarity scores to identify the two classes ('similar' and 'unsimilar'). Different thresholds produce different results in terms of precision, recall, and F1-score when compared to the annotated dataset. A threshold that is too low classifies all sentences as 'similar', and conversely, a value that is too high could lead to classifying all pairs of sentences as 'unsimilar'. The choice of a correct similarity threshold depends on the data under consideration and the specific vector space of a model. Therefore, we optimized its value independently on all models by selecting the best value that maximizes the F1-score on the dataset. The values tested are in the range of 0 to 1 with step 0.001. The best thresholds for the four models are reported in Table 7.

Table 7
Thresholds, F-score (F1), and sentence encoding time of text similarity classification task using ITALIAN BERT and ITALIAN LEGAL BERT variants as bi-encoder.

Encoder	Threshold	F1	Time (secs)
ITA-BERT	0.835	0.7509	96.42
ITA-LEGAL-BERT-FP	0.969	0.7651	96.26
ITA-LEGAL-BERT-SC	0.752	**0.8230**	94.73
DISTIL-ITA-LEGAL-BERT	0.968	0.7868	**18.52**

4.4. Legal document classification

Legal Document Classification is a crucial task in the legal domain that involves assigning specific labels or categories to legal documents based on their content, purpose, or characteristics. This classification process enables efficient organization, retrieval, and analysis of legal documents, facilitating legal research and decision-making. It encompasses identifying the subject matter, type, jurisdiction, or outcome of a legal document.

In this particular context, we are evaluating the performance of ITALIANLEGAL-BERT models on a challenging long text classification task. The objective is to accurately identify the precise legal topic addressed in the document, considering that the documents in question exceed the token limit of 512, which is the capacity of the standard BERT model. To overcome this limitation, we employ the LSG (Local-Sparse-Global) versions of the models, which are specifically designed to handle long texts efficiently.

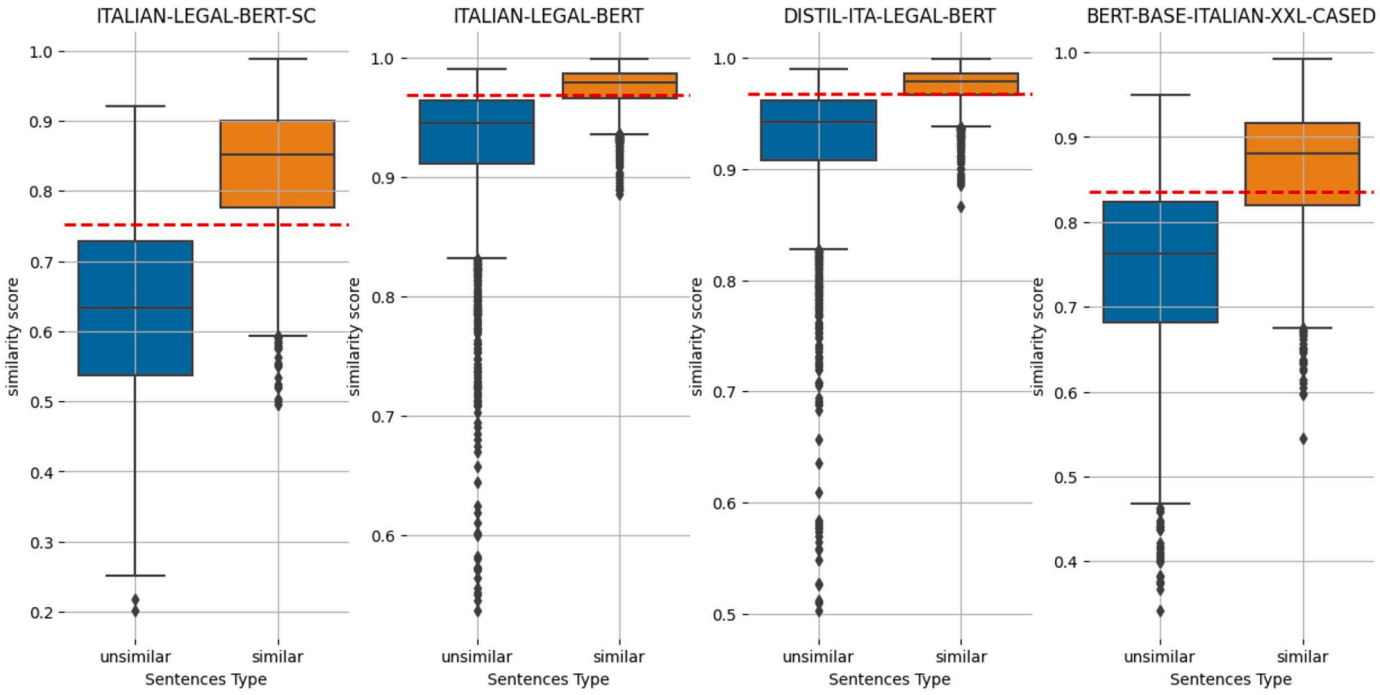

Fig. 2. Box plots showing the different range of semantic similarity scores over the groups of 'similar' and 'unsimilar' pairs of sentences using the Italian BERT and ITALIANLEGAL-BERT models. The red dashed line shows the optimized similarity threshold on the results of the four models.

Table 8
It summarizes the number of tokens using the ITALIAN-LEGAL-BERT tokenizer. It shows the central tendency, dispersion, minimum, maximum, and 25th, 50th, 75th percentiles.

Label	Count	Mean	STD	Min	25%	50%	75%	Max
Joint divorce – Religious	200	1811.5	941	934	1342.2	1558	1914.7	9898
Joint divorce – Civil	200	1760.5	875.1	835	1319.5	1569	1917	9522
Litigation divorce - Religious	200	2021.8	1363.1	733	1186	1514	2216	9606
Litigation divorce - Civil	200	1925.4	1490.5	631	1075	1389	2129.2	10062
Judicial separation	200	2499.1	1947.4	641	945.7	1961	3383.5	11396
TOTAL	1000	2003.7	1402.4	631	1200.7	1561	2195.75	11396

The Table 8 presents a summary of token counts using the ITALIANLEGAL-BERT tokenizer for different labels or categories. It includes statistical measures such as count, mean, standard deviation, minimum, maximum, and percentiles (25th, 50th, and 75th) to describe the distribution of token counts within each label. Additionally, the table provides an overall "TOTAL" row that showcases statistics for the entire dataset. This information offers valuable insights into the length and variability of token counts, providing a better understanding of the characteristics of the legal documents that are being classified.

The evaluation encompasses both the standard ITALIAN-LEGAL-BERT models and their LSG variants, which incorporate the LSG attention mechanism to effectively process long sequences. By comparing the performance of these models, we aim to assess their effectiveness in handling lengthy legal texts and their potential application in real-world scenarios where accurate classification of lengthy documents is required.

The results obtained from this benchmark will provide valuable insights into the capabilities of ITALIAN-LEGAL-BERT models in handling long legal texts and their potential applicability in various legal domains. Moreover, it highlights the advantages of utilizing the LSG versions of the models, which enable improved performance on long text classification tasks that surpass the token limit of the standard BERT model.

Dataset. Also, for this task, we created an ad-hoc dataset by taking 1000 judgments of 5 topics on family law from the Civil Court of Genoa with which we have an agreement. The dataset contains an equal number of documents relating to Civil and Religious joint divorces, Civil and Religious litigation divorces, and separations. It was split stratified on the topic of the document, 70% of the data was used to train the models, 10% of the training data for validation, and 30% for testing.

Training procedure. The training procedure involved fine-tuning the Italian BERT and ITALIAN-LEGAL BERT models, including their LSG (Local-Sparse-Global) variants, with a linear sequence classifier on top. All models were trained using the same hyperparameter configuration, as detailed in Table A.15 in the Appendix. To handle the challenge of processing long documents efficiently, we set the maximum sequence length for LSG models to 2500 tokens, which corresponds to the 80th percentile of the dataset. This choice of sequence length aims to balance capturing the necessary information while maintaining computational efficiency. In addition to the maximum sequence length, we also explored a sliding window approach to handle very long documents. This approach involves dividing a long text into smaller chunks, typically consisting of 512 tokens, with a certain amount of overlap between the chunks. In this case, we used a 50 percent overlap to minimize information loss that may occur due to abrupt cuts. Each chunk was then assigned a document label.

During training, the model was trained on these sub-sequences (or chunks) to predict the corresponding document label. The final prediction for a document was determined by selecting the mode (most frequent prediction) across all sub-sequences (or chunks) of the document. This aggregation strategy helps consolidate predictions and provide a single label for the entire document. By employing this training procedure, we aim to optimize the models' performance in accurately classifying long legal documents, effectively utilizing the strengths of the LSG variants and handling document length variations through the sliding window approach.

Table 9
MCC scores and train time in seconds on document classification datasets.

Model	Max seq	sliding	MCC (%)	Runtime (s)
ITA-BERT	512	FALSE	85.05	153
	512	TRUE	75.08	468
ITA-LEGAL-BERT-FP	512	FALSE	85.85	161
	512	TRUE	85.91	787
LSG- ITA-LEGAL-BERT-FP	2500	FALSE	82.97	705
	2500	TRUE	87.56	1053
ITA-LEGAL-BERT-SC	512	FALSE	86.29	151
	512	TRUE	81.62	400
LSG-ITA-LEGAL-BERT-SC	2500	FALSE	87.94	716
	2500	TRUE	**91.09**	965
DISTIL-ITA-LEGAL-BERT	512	FALSE	85.44	**73**
	512	TRUE	74.10	134

Evaluation. In the evaluation Table 9, it was observed that the LSG version of the ITALIAN-LEGAL-BERT model, which was pre-trained from scratch, achieved better performance (the highest MCC score of 91.09%) compared to other models. This indicates that the LSG variant effectively captures the complex patterns and dependencies present in long legal documents, leading to more accurate predictions. Both the sliding window approach and the non-sliding window approach were evaluated, and the LSG model consistently outperformed the other models in both scenarios. This suggests that the LSG variant not only excels in handling long documents but also maintains its performance when utilizing the sliding window technique to handle document length limitations (Table 9).

Additionally, the distilled model, which incorporates knowledge distillation techniques, showed superior performance compared to the model trained with general knowledge. Notably, the distilled model exhibited more than twice the training speed, indicating its efficiency in training without compromising performance (MCC score of 85.44% without sliding).

Overall, the evaluation results highlight the effectiveness of the LSG version of the ITALIAN-LEGAL-BERT model for document classification tasks, its compatibility with the sliding window approach, and the advantages of utilizing knowledge distillation techniques for improved performance and training efficiency.

5. Limitations

The main limitations come from the limited computational resources with which our models were trained. We are aware that a larger batch size, extended parameters optimization, and a larger data set could lead to better results. Secondly, the type of downstream task could be a limiting factor in model performance. ITALIAN LEGAL BERTs are designed to improve current performance in complex Italian legal tasks, where domain knowledge is very important.

Notably, there exist scenarios where domain-adaptive pre-training, as exemplified by LEGAL-BERT, may not yield a substantial advantage over finetuning. Such circumstances have been explored in studies that scrutinize the nuanced interplay between pre-training and fine-tuning processes in natural language understanding tasks [9]. For instance, certain legal prediction tasks may exhibit intricacies where fine-tuning,

with its ability to adapt more directly to task-specific data, could prove to be equally or even more effective than pre-training on large-scale legal corpora. Moreover, it is imperative to recognize that domain-adaptive pre-training, particularly on extensive legal datasets, entails more demanding requirements in terms of training time, memory usage, and computational resources compared to fine-tuning. The decision to opt for pre-training should, therefore, be driven not only by the task but also by practical constraints and available resources [37]. In addition, all downstream tasks on which the models were tested used civil law data. The main motivation lies in the project on which this work is based, namely the optimization of the civil justice system through the use of artificial intelligence techniques (predictivejustice.eu). For a complete evaluation on Italian law, the models should be evaluated downstream of different types of jurisdictions (e.g., civil, criminal, and administrative), which we intend to do in the near future.

Another common limitation of all Deep Learning systems is that they are not easily interpreted and maintain biases in the data on which it was trained. In particular, biases in the data can lead the model to generate stereotypical or biased content. The biases analysis depends on the specific downstream task and deserves further investigation. Finally, Legal BERT models may not fully understand the nuances and complexities of the legal system and may provide incorrect or incomplete information. It is important that these models are carefully designed and tested to ensure their reliability and accuracy both from the perspective of the quality of the training dataset and subsequent monitoring to prevent data drift. Overall, Legal Bert models have the potential to greatly improve the efficiency and effectiveness of legal systems. Although there are some legitimate concerns about their use, in many applications for decision support or automation of simple legal tasks, the benefits of these models outweigh the risks.

6. Conclusion and future direction

This article introduces ITALIAN-LEGAL-BERT models, specifically designed to enhance the performance of natural language processing (NLP) tasks in the Italian legal domain. These models provide pre-trained linguistic representations tailored to Italian law, obtained through domain-adapted pre-training (ITALIAN-LEGAL-BERT-FP) or training from scratch on the Italian legal domain (ITALIAN-LEGAL-BERT-SC). Two additional variants have been proposed and evaluated to address specific requirements. The first variant is a compact version of ITALIAN-LEGAL-BERT-FP, offering nearly three times faster inference and training while maintaining a comparable level of performance. This variant is ideal for applications with limited computational resources or scenarios involving large-scale data processing. The second variant leverages the Local-Sparse-Global (LSG) attention mechanism, enabling ITALIAN-LEGAL-BERT models to handle exceptionally long sequences, up to 16 K tokens. Through this approach, improvements have been observed in various downstream tasks within the Italian legal domain, including named entity recognition, sequence classification, sentence similarity, and document classification. The selection of the most suitable model depends on the specific use case, with the distilled model striking a balance between performance and inference/training speed. On the other hand, models trained from scratch demonstrate remarkable focus on the legal domain and excel in highly domain-specific and complex applications. Looking ahead, future work aims to explore the potential of ITALIAN-LEGAL-BERT models in more complex tasks such as rhetorical role identification [38], similar case retrieval, legal text summarization, comprehension, law holdings extraction, and legal question answering. These advancements hold promise for further advancements in NLP legal research, computational law, and legal technology applications, ultimately benefiting the legal community and facilitating more sophisticated legal analysis and decision-making processes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Appendix A. Settings and the Hyperparameters

Table A10, Table A11, Table A12, Table A13, Table A14, Table A15

Table A.10
The settings and the hyperparameters for training the MLM ITALIAN-LEGAL-BERT-FP.

Parameter	Values
architectures	BertForMaskedLM
scheduler	AdamW
adam_beta1	0.9
adam_beta2	0.999
adam_epsilon	1.00E-08
initial learning rate	5.00E-05
lr_scheduler_type	linear
num_attention_heads	12
num_hidden_layers	12
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
intermediate_size	3072
layer_norm_eps	1.00E-12
max_position_embeddings	512
position_embedding_type	absolute
num_train_epochs	4
batch_size	10
vocab_size	32,102
type_vocab_size	2

Table A.11
The settings and the hyperparameters for training the MLM ITALIAN-LEGAL-BERT-SC.

Parameter	Values
adam_beta2	0.999
adam_epsilon	1.00E-08
initial learning rate	2.00E-05
lr_scheduler_type	linear
num_attention_heads	1.20E+01
num_hidden_layers	1.20E+01
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
intermediate_size	3072
layer_norm_eps	0.00001
max_position_embeddings	514
position_embedding_type	absolute
num_train_epochs	4.00E+00
batch_size	10
vocab_size	32,005
type_vocab_size	1

Table A.12
The settings and the hyperparameters for training the DISTIL-ITALIAN-LEGAL-BERT.

Parameter	Values
teacher model	ITALIAN-LEGAL-BERT-FP
student	CamemBERTForMaskedLM
pooling mode	Mean tokens
loss	MSELoss
scheduler	AdamW
adambeta1	0.9
adambeta2	0.999
adam epsilon	1e-06
initial learning rate	1e-4
lr scheduler type	linear
num attention heads	12
num hidden layers	4
hidden act	gelu
hidden dropout prob	0.1
hidden size	768
intermediate size	3072
layer norm eps	1e-12
max position embeddings	512
position embedding type	absolute
num train epochs	4
batch size	24
vocab size	32,102
type vocab size	2
warm-up steps	1000
evaluation step	5000
automatic mixed precision	True

Table A.13
The settings and the hyperparameters for training the pipeline's named entity recognizer.

Parameter	Values
pipeline	["transformer","ner"]
ner architectures	"spacy.TransitionBasedParser.v2"
transformer architectures	Italian-BERT or ITALIAN-LEGAL-BERTs
tokenizer name	BertTokenizer or ITALIAN-LEGAL-BERT-SC
scheduler	AdamW
adam epsilon	1e-07
batch size	128
max steps	20,000
learning rate	5e-05 (linearly lr decay)
num warmup steps	250 steps

(*continued on next page*)

Table A.13 (*continued*)

Parameter	Values
patience early stopping	1600 steps
dropout rate	0.1
accumulate gradient	3
test size	0.2
validation size	0.1 (on best F1 score)

Table A.14
Hyperparameters configuration for sequence classification models.

Parameter	Values
Parameter	Values
tokenizer_name	BertTokenizer or ITALIAN-LEGAL-BERT-SC
transformer architectures	Italian-BERT or ITALIAN-LEGAL-BERTs
classifier	Linear, MLP, LSTM, or CNN
scheduler	AdamW
adam_epsilon	2.00E-05
batch_size	8
Epochs	[1,5]
learning_rate	2.00E-05
num_warmup_steps	0.06
weight_decay	0.01
evaluation_strategy	epoch
test_size	0.1
validation_size	0.1 (on best MCC score)

Table A.15
Hyperparameters configuration for document classification models.

Parameter	Values
tokenizer name	BertTokenizer or ITALIAN-LEGAL-BERT-SC
transformer architectures	Italian-BERT or (LSG-)ITALIAN-LEGAL-BERTs
classifier	Linear
scheduler	AdamW
adam epsilon	1e-08
batch size	8 (4 for LSG Models)
max sequence length	512 (2500 for LSG Models)
Epochs	[1,10]
learning rate	4e-05 (linear schedule)
num warmup steps	0.06
evaluation strategy	500 steps
patience early stopping	3 evaluations
test size	0.3
validation size	0.1 (on best MCC score)

References

[1] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019. https://doi.org/10.1093/bioinformatics/btz682. btz682ArXiv:1901.08746 [cs].

[2] Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8. https://doi.org/10.18653/v1/W19-1909. URL, https://aclanthology.org/W19-1909.

[3] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3615–20. https://doi.org/10.18653/v1/D19-1371. URL, https://aclanthology.org/D19-1371.

[4] Caselli T, Basile V, Mitrović J, Granitzer M. HateBERT: retraining BERT for abusive language detection in English. In: Proceedings of the 5th workshop on online abuse and harms (WOAH 2021). Association for Computational Linguistics; 2021. p. 17–25. https://doi.org/10.18653/v1/2021.woah-1.3. OnlineURL, https://aclanthology.org/2021.woah-1.3.

[5] Polignano M, Basile P, Degemmis M, Semeraro G, Basile V. Alberto: Italian Bert language understanding model for NLP challenging tasks based on tweets. CLiC-it. 2019.

[6] Carofiglio G. Con parole precise. Laterza, Roma Bari: Breviario di Scrittura Civile; 2015.

[7] M. Rosati, Forte e chiaro: Il linguaggio del giudice, IL LINGUAGGIO DEL PROCESSO (2016) 115–9. URL https://www.uniba.it/ricerca/dipartimenti/sistemi-giuridici-ed-economici/e.

[8] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: the muppets straight out of law school. Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics; 2020. p. 2898–904. https://doi.org/10.18653/v1/2020.findings-emnlp.261. OnlineURL, https://aclanthology.org/2020.findings-emnlp.261.

[9] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A. Transformers: state-of-the-Art Natural Language Processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Association for Computational Linguistics; 2020. p. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6. OnlineURL, https://aclanthology.org/2020.emnlp-demos.6.

[10] L. Zheng, N. Guha, B.R. Anderson, P. Henderson, D.E. Ho, When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset, arXiv:2104.08671 [cs] (2021).

[11] Lippi M, Pa lka P, Contissa G, Lagioia F, Micklitz H-W, Sartor G, Torroni P. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of

service. Artif Intell Law 2019;27(2):117–39. https://doi.org/10.1007/s10506-019-09243-2. doi:10.1007/s10506-019-09243-2URL.
[12] G. Zhang, D. Lillis, P. Nulty, Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers 10.
[13] Zhang G, Nulty P, Lillis D. Enhancing legal argument mining with domain pre-training and neural networks. J Data Min Digit Human 2022;NLP4DH:9147. https://doi.org/10.46298/jdmdh.9147. URL, https://jdmdh.episciences.org/9147.
[14] C. Condevaux, S. Harispe, L.S.G. Attention: Extrapolation of pretrained Transformers to long sequences, arXiv:2210.15497 [cs] (Oct. 2022). doi:10.48550/arXiv.2210.15497.
[15] Harold JS, Lee E, Jeffrey AS, Andrew DM, Theodore JR, Sara CB. Supreme court database. Version 2020 release 01. Washington University Law; 2020. URL, http://Supremecourtdatabase.org.
[16] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317–23. https://doi.org/10.18653/v1/P19-1424. URL, https://aclanthology.org/P19-1424.
[17] Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P. Paragraph-level rationale extraction through regularization: a case study on European Court of Human Rights Cases. In: Proceedings of the 2021 conference of the North American Chapter of the Association for computational linguistics: human language technologies. Association for Computational Linguistics; 2021. p. 226–41. https://doi.org/10.18653/v1/2021.naacl-main.22. OnlineURL, https://aclanthology.org/2021.naacl-main.22.
[18] J. Cui, X. Shen, F. Nie, Z. Wang, J. Wang, Y. Chen, A survey on legal judgment prediction: datasets, metrics, models and challenges, arXiv:2204.04859 [cs] (2022). doi:10.48550/arXiv.2204.04859.
[19] M. Masala, R. Iacob, A.S. Uban, M.-A. Cidotâ, H. Velicu, T. Rebedea, M. Popescu, jurBERT: a Romanian BERT model for legal judgement prediction, NLLP (2021). doi:10.18653/v1/2021.nllp-1.8.
[20] Douka S, Abdine H, Vazirgiannis M, Hamdani RE, Amariles DR. JuriBERT: a masked-language model adaptation for French legal text. NLLP; 2021. https://doi.org/10.18653/v1/2021.nllp-1.9.
[21] Xiao C, Hu X, Liu Z, Tu C, Sun M. Lawformer: a pre-trained language model for Chinese legal long documents. AI Open 2021;2:79–84. https://doi.org/10.1016/j.aiopen.2021.06.003. URL, https://www.sciencedirect.com/science/article/pii/S2666651021000176.
[22] M. AL-Qurishi, S. AlQaseemi, R. Soussi, AraLegal-BERT: a pretrained language model for Arabic Legal text, arXiv:2210.08284 [cs] (Oct. 2022). doi:10.48550/arXiv.2210.08284.
[23] Tagarelli A, Simeri A. Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code. Artificial Intelligence and Law 2022;30(3):417–73. https://doi.org/10.1007/s10506-021-09301-8. 10.1007/s10506-021-09301-8. URL.
[24] Licari D, Comand`e G, ITALIAN-LEGAL-BERT: A. Pre-trained transformer language model for Italian Law. In: Symeonidou D, Yu R, Ceolin D, Poveda-Villalón M, Audrito D, Caro LD, Grasso F, Nai R, Sulis E, Ekaputra FJ, Kutz O, Troquard N, editors. Companion Proceedings of the 23rd international conference on knowledge engineering and knowledge management, Vol. 3256 of CEUR workshop proceedings. Bozen-Bolzano, Italy: CEUR; 2022. iSSN: 1613-0073. URL, https://ceur-ws.org/Vol-3256/km4law3.
[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 [cs] 2019). doi:10.48550/arXiv.1810.04805.
[26] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. CoRR 2017. abs/1706.03762arXiv:1706.03762.
[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A. Robustly optimized BERT pretraining approach, arXiv:1907.11692 [cs] (Jul. 2019). doi:10.48550/arXiv.1907.11692.
[28] Schuster M, Nakajima K, Japanese and Korean voice search, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). In: Conference Name: ICASSP 2012 2012 IEEE international conference on acoustics, speech and signal processing. Kyoto: Japan Publisher: IEEE; 2012. p. 5149–52. https://doi.org/10.1109/ICASSP.2012.6289079. ISBN: 9781467300469 9781467300452 9781467300445 PlaceURL, http://ieeexplore.ieee.org/document/6289079/.
[29] Martin L, Muller B, Suarez PJO, Dupont Y, Romary L, de la Clergerie EV, Seddah D, Sagot B. CamemBERT: a tasty French Language Model. In: Proceedings of the 58th annual meeting of the association for computational linguistics; 2020. p. 7203–19. https://doi.org/10.18653/v1/2020.acl-main.645. arXiv:1911.03894 [cs].
[30] T. Kudo, J. Richardson, SentencePiece: a simple and language independent subword Tokenizer and Detokenizer for Neural Text Processing, arXiv:1808.06226 [cs] (Aug. 2018). doi:10.48550/arXiv.1808.06226.
[31] Mattmann CA, Zitting JL. Tika in action. Shelter Island, NY: Manning Publications; 2012. oCLC: ocn731912756.
[32] Reimers N, Gurevych I. Sentence-Bert: sentence embeddings using Siamese Bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2019. URL, http://arxiv.org/abs/1908.10084.
[33] M. Honnibal, I. Montani, spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear (2017).
[34] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: text annotation tool for human, software available from https://github.com/doccano/doccano (2018). URL https://github.com/doccano/doccano.
[35] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. CoRR 2016. abs/1603.01360.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Duchesnay, Scikitlearn: machine learning in Python, arXiv:1201.0490 [cs] (2018). doi:10.48550/arXiv.1201.0490.
[37] Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 3645–50. https://doi.org/10.18653/v1/P19-1355. URL, https://aclanthology.org/P19-1355.
[38] Walker VR, Pillaipakkamnatt K, Davidson AM, Linares M, Pesce DJ. Automatic classification of rhetorical roles for sentences: comparing rule-based scripts with machine learning. ASAIL@ICAIL. 2019.