ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law

Daniele Licari[1,*], Giovanni Comandè[1]

[1]EMbeDS, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy.

Abstract

The state of the art in natural language processing is based on transformer models that are pre-trained on general knowledge and enable efficient transfer learning in a wide variety of downstream tasks even with limited data sets. However, these models significantly decrease performance when operating in specific and sectoral domains. This is problematic in the Italian legal context, as there are many discrepancies between the language found in generic open source corpora (e.g., Wikipedia and news articles) and legal language, which can be cryptic, Latin-based, and domain idiolectal formulas.

In this paper, we introduce the ITALIAN-LEGAL-BERT model with additional pre-training of the Italian BERT model on Italian civil law corpora. It achieves better results than the 'general-purpose' Italian BERT in different domain-specific tasks.

Keywords

Legal artificial intelligence, Pre-trained language model, Italian Legal BERT

1. Introduction

In many domains, specialized models performed better than pre-trained models on general domains[1, 2, 3, 4, 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 in complex tasks.

In the Italian legal context, the discrepancy between specific language and 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 useless stylistic expedients that often forcibly show a continuity with the languages of the past (Latin or old Italian). The full understanding of judicial texts is the exclusive prerogative of domain experts. It contains technicalities with specific and unambiguous meanings ("contumacia", "anticresi", "anatocismo", "sinallagma"). It also makes extensive use of terms in general use but often employed with their own and specific meanings, if not entirely different from those in common use. For example, "nullità", "annullabilità", "inefficacia", "inutilizzabilità", which outside of legal language are synonyms of annulment, denote entirely distinct and different 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 [6].

EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, September 26–29, 2022, Bozen-Bolzano, IT

*Corresponding author.

d.licari@santannapisa.it (D. Licari); g.comande@santannapisa.it (G. Comandè)

Chalkidis et al. [7] developed the first transformers-based model for the English legal domain (LEGAL-BERT) by improving the performance of the general-purpose model (BERT-BASE) in several prediction tasks. The basic idea is that a model with legal domain knowledge can classify legal documents better than a model with general knowledge.

Taking inspiration from LEGAL-BERT, we report on the development of the ITALIAN-LEGAL-BERT model capable of understanding the semantic meaning of Italian legal texts by additional pre-training of ITALIAN XXL BERT (available on hunggingface hub[8]) on Italian civil law corpora.

In this work, we make the following contributions:

1. We publicly release[1] ITALIAN-LEGAL-BERT to assist Italian legal NLP research. It is, to the best of our knowledge, the first pre-trained language model further trained on a large corpus of Italian civil cases.
2. We demonstrate that ITALIAN-LEGAL-BERT outperforms the generalized equivalent in terms of perplexity (PPT) and end results in downstream tasks such as sequence classification, semantic similarity, and named entities recognition in the Italian legal domain.
3. We also evaluated the model on anonymized datasets to explore whether it is biased toward demographic information and personal data.

2. Related Work

The legal writing system differs greatly from generic texts with many domain-specific peculiarities. Some researchers demonstrated that the use of domain-specific pre-trained models can improve the performance of downstream tasks in the legal domain.

Chalkidis et al. [7] proposed the LEGAL-BERT model pre-trained from scratch on 11.5 GB of legal texts and its variant further pretraining BERT-base on legal corpora. Their experiments indicated more substantial improvement in the most challenging end-task (i.e., multi-label classification in ECHR-CASES and contract header, lease details in CONTRACTS-NER) where in-domain knowledge is more important. In addition, no significant differences were found in performance between the two LEGAL-BERT variants.

Similar evidence is reported by Zheng et al. [9]. 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 different downstream NLP tasks with different difficulties and domain specificity. They suggest using domain-specific pre-trained models for highly difficult legal tasks. They performed better than BERT-base in complex downstream tasks such as identifying whether contract terms are potentially unfair [10]. In contrast, additional domain pretraining adds little value to simpler tasks compared to BERT.

The recent works of Zhang et al [11, 12] 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.

[1] on huggingface.co/dlicari/Italian-Legal-BERT

Success in this area encouraged researchers to create pre-trained language models on legal corpora in different languages [13]. Masala et al 2021[14] released the jurBERT model pre-trained on a large Romanian legal corpus. It outperformed several strong baselines for legal judgment prediction. In the same year, Douka et al [15] created a language model adapted to French legal text demonstrating that their model works betters in the French legal domain than their generalized equivalents. In China, researchers [16] have improved many predictive tasks on long Chinese legal documents through a pre-trained language model on millions of documents published by the Chinese government.

In Italy, A. Tagarelli and A. Simeri 2022[17] 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. Unfortunately, they did not provide a direct comparison with the Italian BERT model on which the domain adaptation was performed. Therefore, it was not possible to evaluate the advantages of domain fitting of the BERT model over the equivalent generalized model in the reported downstream tasks.

The work cited above differs from ours in terms of the reference corpus, problems addressed, and analysis of results. First, our model was trained using a large amount of decrees, ordinances, and judgments of Italian courts. They may include, in addition to the cited laws of the civil code, judge's reasons, facts, decisions, reasons, proposals of the parties, medico-legal information, legal rules, verified evidence, witnesses, etc. Second, this variety of information and the size of the training dataset allowed us to create a language model that better represents the Italian legal context by capturing the complex semantic interactions between facts, reasons, and laws. Therefore, our model can be applied to more complex general tasks, such as identifying rhetorical roles, retrieving similar cases, extracting arguments, argument mining, legal reading comprehension, and legal question answering. Third, our analysis focused on directly comparing the generalized Italian BERT model and the adapted model on the legal domain ITALIAN-LEGAL-BERT, to assess the improvements achieved in several downstream tasks. Finally, our model was shared on the Huggingface platform, to maximize usability and make a concrete contribution to the growth of NLP applications in the Italian legal context.

3. Italian Legal BERT

Background. BERT (Bidirectional Encoder Representations from Transformers [18]) is a contextual word embedding model using the transformers architecture [19] 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.

Italian BERT. The Italian XXL BERT model (cased, 12-layer, 768-hidden, 12-heads, 110M parameters) has the Bidirectional Encoder Representations from Transformers architecture and has been trained on large Italian corpora 81 GB derived from Wikipedia in Italian, various texts from the OPUS corpora collection (opus.nlpl.eu), and data from the Italian part of the OSCAR corpus (oscar-corpus.com). It is available on the Huggingface model hub [8] and trained by the

Table 1
Italian Legal BERT and Italian XXL BERT Perplexity scores on evaluation datasets. Lower perplexity indicates better performance

Eval Dataset	N. Documents	N. Sentences	Model	Perplexity
Civil cases (pst.giustizia.it)	20,000	566,000	ITALIAN-LEGAL-BERT	**8.9892**
			Italian XXL BERT	10.9891
Criminal cases (italgiureweb)	21,000	702,677	ITALIAN-LEGAL-BERT	**5.0518**
			Italian XXL BERT	6.0515

MDZ Digital Library team at the Bavarian State Library.

Training procedure. We initialized ITALIAN-LEGAL-BERT with ITALIAN XXL BERT and pretrained for an additional 4 epochs on 3.7 GB of text from the National Jurisprudential Archive using the Huggingface PyTorch-Transformers library [8]. We used BERT architecture with a language modeling head on top, AdamW Optimizer, initial learning rate 5e-5 (with linear learning rate decay, ends at 2.525e-9), sequence length 512, batch size 10 (imposed by GPU capacity), 8.4 million training steps, device 1*GPU V100 16GB. More details on the hyperparameters we consider for each training phase can be found in the appendix.

Training Dataset. National Jurisprudential Archive (Archivio Giurisprudenziale Nazionale, pst.giustizia.it) is a public repository containing millions of legal documents (decrees, orders, and civil judgments) from Italian courts and courts of appeal. We downloaded about 235,000 documents as PDF files. The documents were converted to plain text using the Tika framework [20].

Preprocessing Dataset. We preprocessed the case law corpus with some cleaning functions. We compacted whitespace and new lines using a regular expression. The sentence segmentation process was customized by adding new tokenization rules to the spaCy model for the Italian language. The added exceptions concern abbreviations and acronyms used in Italian legal texts[2]. Segmented sentences were cleaned up by removing all special characters through an additional expression rule. The final corpus contains 21,004,500 sentences and 498,002,402 words (3.7 GB). The final model input was created using the Italian BERT tokenizer on the corpus sentences, truncating them to the maximum length (512 tokens).

Evaluation Dataset. We downloaded an additional 20,000 civil cases from the National Jurisprudential Archive. We applied the same preprocessing procedure as the training set to create a corpus containing 566,000 sentences and 17,936,466 words. In order to evaluate the performance in the criminal context we have also downloaded 21,000 criminal cases from italgiureweb (italgiure.giustizia.it) corpus containing 702,677 sentences and 20,164,194 words. Finally, we applied random masking (15% tokens) to the sentences in both datasets.

MLM Evaluation. Perplexity (PPL) is one of the most common metrics for evaluating language models. It is the exponential of the cross-entropy loss, a lower perplexity indicates a better model. The perplexity for the MLM objective is computed to make predictions for the masked tokens (which represent 15% of the total here) while having access to the rest of the tokens.

[2]The full list is available at https://huggingface.co/dlicari/Italian-Legal-BERT/blob/main/abbreviazioni.csv

The results in Table 1 showed that ITALIAN-LEGAL-BERT dropped perplexity by 18.2% in civil cases and by 15.4% in criminal cases with respect to Italian XXL BERT. Lower perplexity scores on criminal cases could indicate greater use of commonly used notions than on civil cases.

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 BERT**
Il ~~**padre**~~ può vedere il figlio a weekend alternati en: The ~~**father**~~ can see his son on alternate weekends	1. 'genitore' (53.61%) 2. **'padre'** (27.70%) 3. 'papà' (6.81%) 4. 'marito' (2.19%) 5. 'proprietario' (0.62%)	1. **'padre'** (99.24%) 2. 'genitore' (0.56%) 3. 'ricorrente' (0.05%) 4. 'resistente' (0.03%) 5. 'papà' (0.03%)
viene ~~**stabilita**~~ una collocazione paritetica dei figli en: an equal placement of the children is ~~**established**~~.	1. 'garantita' (24.1%) 2. 'meno' (10.72%) 3. 'proposta' (6.66%) 4. **'stabilita'** (4.52%) 5. 'assicurata' (4.09%)	1. 'prevista' (40.48%) 2. **'stabilita'** (21.81%) 3. 'disposta' (12.74%) 4. 'assicurata' (6.32%) 5. 'garantita' (1.77%)
assegno di mantenimento comprensivo di spese ~~**straordinarie**~~ en: maintenance allowance including ~~**extraordinary**~~ expenses.	1. '.' (38.58%) 2. 'mediche' (17.01%) 3. ';' (6.62%) 4. 'legali' (4.55%) 5. 'generali' (3.35%)	1. **'straordinarie'** (69.25%) 2. ':' (7.61%) 3. 'extra' (4.86%) 4. 'mediche' (4.30%) 5. '.' (4.20%)
viene stabilito il ~~**mantenimento**~~ diretto en: direct ~~**maintenance**~~ is established	1. 'trattamento' (8.58%) 2. 'prezzo' (7.43%) 3. 'contratto' (5.08%) 4. 'contributo' (4.23%) 5. 'lavoro' (4.06%)	1. 'pagamento' (48.93%) 2. 'versamento' (23.89%) 3. **'mantenimento' (5.20%)** 4. 'trasferimento' (2.54%) 5. 'rimborso' (2.09%)
cambiamento di ~~**sesso**~~ senza operazione chirurgica en: ~~**sex**~~ change without surgery	1. 'pelle' (19.12%) 2. 'capelli' (16.54%) 3. **'sesso'** (8.53%) 4. 'colore' (6.17%) 5. 'peso' (4.48%)	1. **'sesso'** (89.01%) 2. 'genere' (5.40%) 3. 'nome' (1.20%) 4. 'profilo' (1.01%) 5. 'persona' (0.43%)
Il ~~**ricorrente**~~ ha chiesto revocarsi l'obbligo di pagamento. en: The ~~**plaintiff**~~ requested that the payment obligation be revoked.	1. 'Comune' (11.89%) 2. 'giudice' (9.17%) 3. 'cittadino' (4.70%) 4. 'lavoratore' (3.17%) 5. 'sindaco' (2.62%)	1. **'ricorrente'** (72.64%) 2. 'convenuto' (9.64%) 3. 'resistente' (3.99%) 4. 'lavoratore' (2.90%) 5. 'Ministero' (2.53%)
Non avendo la ~~**Corte**~~ di merito valutato la prova en: Not having the ~~**Court**~~ of merit assessed the evidence	1. 'Commissione' (41.46%) 2. 'commissione' (21.54%) 3. 'classe' (18.65%) 'valutazione' (6.84%) 5. 'prova' (3.02%)	1. **'Corte'** (56.29%) 2. 'corte' (23.26%) 3. 'sentenza' (12.49%) 4. 'giurisprudenza' (3.87%) 5. 'decisione' (1.64%)

Fill Mask. A further qualitative investigation was conducted by asking the judges for some domain sentences and making an inference about a mask word contained in the sentence. We

used the mask filling pipeline of the Hugging Face Transformers library to return the top 5 suggestions for the masked word. Tab 2 reports the results on Italian BERT and ITALIAN-LEGAL- BERT models, the strikethrough words have been masked to be predicted by the models.

This analysis helps us to better study the implicit knowledge that the ITALIAN-LEGAL-BERT model has accumulated during pre-training. As can be seen in Table 2, the correct word always appears in the top three in the inference made with the ITALIAN-LEGAL-BERT and indicates that our model succeeds in capturing the specific context better than the general model.

4. Downstream evaluation task

The Italian BERT and ITALIAN-LEGAL-BERT models were evaluated and compared on three domain-specific downstream tasks. In the first task, we trained the models with an additional sequence tagging layer on the top using spaCy[21] to recognize name/role of actors involved in the trial. For the second task, we trained the models with a sequence classification head (a linear layer on top of the pooled output) for the classification of sentence type. In the last 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.

4.1. Named Entity Recognition

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

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

Entity	Description	Train	Test
Person	Names of the main actors and defendants	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

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 significant experience of our research group on these issues, the selected judgments are all those of personal injury (No. 59) contained in the database, and an equal number of family judgments was selected stratified to the text's length. Next, we converted the PDF files to plain text using Tika [20], applied some text cleaning functions (removal of multiple blank lines and extra spaces), and converted the texts to an annotable data structure (jsonl format) to import them into the Doccano annotation tool [22]. 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 6,355 entities; Table 3 shows the distribution of entities on the dataset. Finally, the dataset was split 80% for model training (10% of the training set for validation) and 20% for model evaluation in a stratified 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 [21]. The model architecture consists of a two-tier pipeline: the contextual embedding layer and the transition-based chunking model [23]. The first uses pre-trained language models to encode tokens into continuous vectors based on their context. The second predicts text structure by mapping it onto a set of state transitions. It uses the output (contextual word embeddings) from the previous step to incrementally construct states from the input sequence and assign them an entity label using a multilayer neural network. We trained and compared two spacy-based entity recognition pipelines using Italian BERT and ITALIAN-LEGAL-BERT as the contextual embedding layer.

Training procedure. We trained two named entity recognition pipelines, Italian BERT + spaCy's NER and ITALIAN-LEGAL-BERT+ spaCy's NER, using AdamW Optimizer, initial learning rate 5e-5 (with linearly decay), 20000 maximum number of steps, 250 warm-up steps, early stopping patience on the F1 validation score, and batch size 128 (see Table 10 in the Appendix for more details).

Evaluation. We compared the two NER pipelines using the exact match criterion with gold-standard entities (both entity boundary and type are correct) in the test set. Precision, recall, and F-score are used to evaluate and compare the performance. The results in Table 4 show that the NER pipeline with ITALIAN-LEGAL-BERT contextual encoder outperforms that with Italian BERT in recognizing most entities.

Table 4
Precision (P), recall (R) and F-score (F1) for The ITALIAN LEGAL BERT+Spacy NER and ITALIAN LEGAL BERT+Spacy NER models evaluated using exact match criterion on individual entities and macro-average on the test set. The Support is the number of samples in the different entity types.

	ITALIAN LEGAL BERT+Spacy NER			**ITALIAN BERT+Spacy NER**			
Type	P	R	F1	P	R	F1	**Support**
Person	76.41	93.74	**84.19**	78.60	89.57	83.73	771
Person-Judge	97.53	98.75	**98.14**	95.24	100	97.56	80
Person-Lawyer	88.46	94.54	**91.39**	85.90	91.78	88.74	74
Person-Expert	73.08	79.17	**76.00**	66.33	79.17	70.37	24
Person-Witness	79.49	55.36	**65.26**	69.23	32.14	43.90	57
Person-Family	41.46	10.56	**16.83**	38.46	6.21	10.70	162
Person-Family-Children	60.00	47.37	52.94	46.41	74.74	**57.26**	95
Macro	73.78	68.50	**69.25**	68.17	67.66	64.61	

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; an indication 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 for the decision (the statement of reasons);
5. CONCLUSION: the decisional content of the judgment.

We want to evaluate the ITALIAN-LEGAL-BERT model 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.

Benchmark Dataset. We randomly selected 6,190 sentences from documents with 5 sections (using regular expression) from Italian Civil Law DB (pst.giustizia.it) stratified on section length, Table 5.

Finally, the dataset was split 80% for training models and 20% for model evaluation in a stratified fashion on the section name to preserve the distribution of sentences across both subsets. The training set was further divided using its 10% for validation.

Table 5
Distribution of sentences over the 5 sections

SECTION NAME	N. SENTENCES
INTRODUCTION	**560**
CONCLUSION OF THE PARTIES	**1,862**
DEVELOPMENT OF THE TRIAL	**949**
REASON	**1,810**
CONCLUSION	**1,009**
TOTAL	**6,190**

Training procedure. We trained Italian BERT and ITALIAN-LEGAL-BERT models with a sequence classification head on top (a linear layer on top of the pooled output) using the same hyperparameters configuration for both (in Table 11 in the Appendix). The final models were trained at the best epoch with a higher validation MCC (Matthew's correlation coefficient) score in the range of 1 to 7 epochs (5 was the best epoch for Italian BERT and 3 for ITALIAN-LEGAL-BERT).

Evaluation. We compared the results on the test set of the two models, Italian BERT and ITALIAN-LEGAL-BERT, trained with the same configuration (Table 11). Models' performance was evaluated on the F1 MACRO and MCC scores on the test set. The results in Table 6 show that the pre-trained model on the Italian legal domain (0.89 F1, 0.83 MCC) outperforms the "general-purpose" models (0.869 F1, 0.806 MCC) in this sentence classification task.

Table 6
F1 and MCC for sentence classification using Italian BERT and ITALIAN-LEGAL-BERT models.

Model	F1	MCC
Italian BERT	0.869	0.806
ITALIAN-LEGAL-BERT	**0.890**	**0.830**

Model Bias. Similar to Chalkidis et al. [24], we investigated how sensitive our model is to personal data. The main information may concern "parties", "witnesses", "important companies", "identifiers", "dates" or "places". The purpose is to understand whether the model over-fits on these data and makes decisions based on demographic and personal information. E.g., 'Mario Rossi' is a judge then it is a 'Decision' sentence, or 'Daniele Licari' is a defendant then it is a 'Conclusion of the parties' sentence. The following experiments focused on the sensitivity of our models to such information by training and evaluating the models on an anonymized version of the dataset.

For entity recognition to be anonymized, we used the model from previous work ([25]) based on pre-trained Transformers embeddings and the transition-based chunking model of spaCy. It found 6,393 entities to be anonymized on the dataset (6,190 sentences). We applied two different anonymization strategies: OMISSIS and TAGGING. The OMISSIS strategy replaced Named Entities with a fixed value (e.g. "Daniele lives in Milan" -> "OMISSIS lives in OMISSIS"). The TAGGING strategy replaced Named Entities with the entity name (e.g. "Daniele lives in Milan" -> "PERSON lives in LOCATION).

The two versions of the anonymized dataset were used to train the two sentence classification models, with Italian BERT and ITALIAN-LEGAL-BERT, using the same configuration and training procedure performed on the raw data. Table 7 shows the comparison of results on the classification models trained on raw and anonymized datasets.

Table 7
F1 and MCC for sentence classification on raw and anonymized using Italian BERT and ITALIAN-LEGAL-BERT models.

Model	Anonymization Strategy	F1	MCC
Italian BERT	NO	0.869	0.806
	OMISSIS	0.861	0.795
	TAGGING	0.862	0.795
Italian Legal BERT	NO	0.890	0.830
	OMISSIS	0.890	0.830
	TAGGING	0.866	0.827

The results of the models on the anonymized dataset and the original dataset are very similar, which might indicate that personal data are not relevant for section prediction.

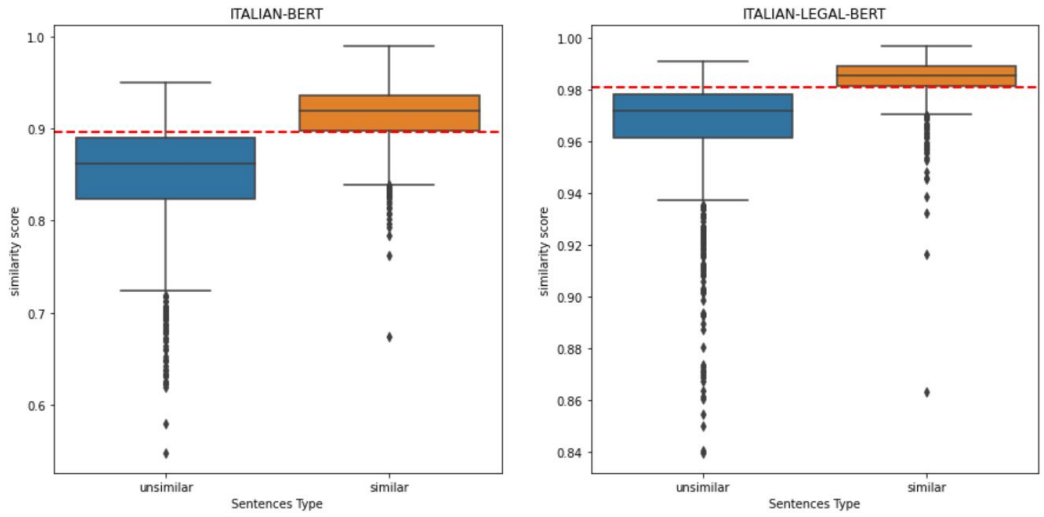
4.3. Semantic Similarity

We tested the ability of the model on the task of determining whether two pieces of 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 a subset of 1,000 judgments from the Italian Civil Law DB, 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 2,000 text pairs (1,000 labeled as 'similar' and the other 1,000 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.

Similarity Procedure. The semantic similarity between the text pairs in the dataset was evaluated using both the Italian BERT and ITALIAN-LEGAL-BERT 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 ($cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$). For each model, a similarity threshold was established to identify similar and non-similar texts. The Figure 1 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.

Figure 1: 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 ITALIAN-LEGAL-BERT models. The red dashed line shows the optimized similarity threshold on the results of the two models.



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 both 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 experiments suggest 0.897 as the best threshold for Italian BERT and 0.981 for ITALIAN-LEGAL-BERT (the red dashed lines in Figure 1).

Evaluation. The performance of text similarity classifications with optimized threshold was evaluated with precision, recall, and F1 score based on true labels. The experimental results, reported in Table 8, show that the ITALIAN-LEGAL-BERT model outperformed the Italian BERT model in this downstream semantic similarity task.

Table 8
Precision (P), recall (R) and F-score (F1) of text similarity classification task using ITALIAN LEGAL BERT and Italian LEGAL

Model	P	R	F1
Italian BERT	0.791	0.789	0.789
ITALIAN-LEGAL-BERT	**0.825**	**0.822**	**0.822**

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.

Another limitation concerns the use of a single data source. Unlike the English language, it is not easy to find large legal corpora on which to train domain-specific models. Although the dataset contains decrees, orders, and judgments from all Italian courts, we did not consider criminal law in our training. However, we have evaluated the perplexity of mask filling on more than 20,000 criminal cases obtaining results similar to the civil context. This suggests to us that the model might work well in the criminal context as well, but further investigation of downstream legal tasks is needed. In addition, although the model was evaluated on a different dataset of the pretraining data, the civil evaluation dataset could still contain some documents written by the judges themselves which could affect the gain of ITALIAN-LEGAL-BERT. We think it is a small gain since the criminal case evaluation dataset (written by different judges) is still significant compared to the generic Italian BERT model.

Moreover, the type of downstream task could be a limiting factor in model performance. ITALIAN LEGAL BERT is designed to improve current performance in complex Italian legal tasks, where domain knowledge is very important. As suggested by experiments on English Legal-BERT[9], using the model in simple downstream tasks may not lead to improvements

over the model trained on general knowledge or even worsen performance.

Finally, a 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. We explore if models are biased towards demographic and personal information via data anonymization, but the analysis depends on the specific downstream task and deserves further investigation.

6. Conclusion and Future Direction

In this article, we introduced ITALIAN-LEGAL-BERT which aims to improve the outcomes of downstream NLP tasks in the Italian legal domain and contribute to the advancement of NLP legal research, computational law, and legal technology applications. It is a pre-trained linguistic representation for Italian law based on ITALIAN BERT XXL with additional pretraining on 235,000 civil cases (domain-adaptive pretraining). We compared the ITALIAN-LEGAL-BERT and Italian BERT models on the downstream tasks of identifying named entities by person type, semantic similarity, and classifying rhetorical sentences by section class. We demonstrated that it can improve the performance of the 'general-purpose' model on downstream tasks in the Italian legal domain. In the future, we plan to exploit the ITALIAN-LEGAL-BERT's potential and test it on more complex tasks, such as rhetorical role identification (e.g. evidence, legal rule, reasoning, decision) [26], similar case retrieval, legal reading comprehension, and legal question answering. In addition, we are working to test it in combination with other deep learning architectures (LSTM, CNN) to achieve better results. Finally, we intend to release new versions of the ITALIAN-LEGAL-BERT pre-trained from scratch on the large Italian legal corpora.

References

[1] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019) btz682. URL: http://arxiv.org/abs/1901.08746. doi:10.1093/bioinformatics/btz682, arXiv:1901.08746 [cs].

[2] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly Available Clinical BERT Embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://aclanthology.org/W19-1909. doi:10.18653/v1/W19-1909.

[3] I. Beltagy, K. Lo, A. Cohan, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://aclanthology.org/D19-1371. doi:10.18653/v1/D19-1371.

[4] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, 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, Online, 2021, pp. 17–25. URL: https://aclanthology.org/2021.woah-1.3. doi:10.18653/v1/2021.woah-1.3.

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

[6] M. Rosati, Forte e chiaro: Il linguaggio del giudice, IL LINGUAGGIO DEL PROCESSO (2016) 115–119. URL: https://www.uniba.it/ricerca/dipartimenti/sistemi-giuridici-ed-economici/edizioni-digitali/i-quaderni/Quaderni62017Triggiani.pdf.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The Muppets straight out of Law School, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, 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, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[9] 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, 2021. URL: http://arxiv.org/abs/2104.08671, arXiv:2104.08671 [cs].

[10] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, P. Torroni, CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service, Artificial Intelligence and Law 27 (2019) 117–139. URL: https://doi.org/10.1007/s10506-019-09243-2. doi:10.1007/s10506-019-09243-2.

[11] 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.

[12] G. Zhang, P. Nulty, D. Lillis, Enhancing Legal Argument Mining with Domain Pre-training and Neural Networks, Journal of Data Mining & Digital Humanities NLP4DH (2022) 9147. URL: https://jdmdh.episciences.org/9147. doi:10.46298/jdmdh.9147.

[13] J. Cui, X. Shen, F. Nie, Z. Wang, J. Wang, Y. Chen, A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges, 2022. URL: http://arxiv.org/abs/2204.04859. doi:10.48550/arXiv.2204.04859, arXiv:2204.04859 [cs].

[14] M. Masala, R. Iacob, A. S. Uban, M.-A. Cidotã, 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.

[15] S. Douka, H. Abdine, M. Vazirgiannis, R. E. Hamdani, D. R. Amariles, JuriBERT: A Masked-Language Model Adaptation for French Legal Text, NLLP (2021). doi:10.18653/v1/2021.nllp-1.9.

[16] C. Xiao, X. Hu, Z. Liu, C. Tu, M. Sun, Lawformer: A pre-trained language model for Chinese legal long documents, AI Open 2 (2021) 79–84. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000176. doi:10.1016/j.aiopen.2021.06.003.

[17] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code, Artificial Intelligence and Law 30 (2022) 417–473. URL: https://doi.org/10.1007/s10506-021-09301-8. doi:10.1007/s10506-021-09301-8.
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805. doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.
[20] C. A. Mattmann, J. L. Zitting, Tika in action, Manning Publications, Shelter Island, NY, 2012. OCLC: ocn731912756.
[21] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.
[22] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.
[23] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, CoRR abs/1603.01360 (2016). URL: http://arxiv.org/abs/1603.01360. arXiv:1603.01360.
[24] I. Chalkidis, I. Androutsopoulos, N. Aletras, Neural Legal Judgment Prediction in English, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4317–4323. URL: https://aclanthology.org/P19-1424. doi:10.18653/v1/P19-1424.
[25] D. Licari, M. F. Romano, G. Comandé, Anonymization of italian legal textual documents using deep learning, volume 2 of *Proceeding of the 16th International Conference on Statistical Analysis of Textual Data (JADT22)*, VADISTAT Press / Edizioni Erranti, Naple, 2022, pp. 552–559.
[26] V. R. Walker, K. Pillaipakkamnatt, A. M. Davidson, M. Linares, D. J. Pesce, Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning, in: ASAIL@ICAIL, 2019.

A. Settings and the Hyperparameters

Table 9
The settings and the hyperparameters for training the MLM ITALIAN-LEGAL-BERT

Parameter	Values
architectures	['BertForMaskedLM']
scheduler	AdamW
adam_beta1	0.9
adam_beta2	0.999
adam_epsilon	1e-08
initial learning rate	5e-5
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	1e-12
max_position_embeddings	512
position_embedding_type	absolute
num_train_epochs	4
batch_size	10
vocab_size	32102
type_vocab_size	2

Table 10
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-BERT"
tokenizer_name	BertTokenizer
scheduler	AdamW
adam_epsilon	1e-07
batch_size	128
max_steps	20000
learning_rate	5e-05 (linearly lr decay)
num warmup steps	250 steps
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 11
Hyperparameters configuration for sequence classification models

Parameter	Values
architectures	['BertForSequenceClassification']
tokenizer_name	BertTokenizer
scheduler	AdamW
adam_epsilon	2e-05
batch_size	8
Epochs	[1,7]
learning_rate	5e-05
num_warmup_steps	0.06
test_size	0.2
validation_size	0.1 (on best MCC score)