PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction

PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction

Irene Benedetto[1,2] **Alkis Koudounas**[1] **Lorenzo Vaiani**[1] **Eliana Pastor**[1]

Elena Baralis[1] **Luca Cagliero**[1] **Francesco Tarasconi**[2]

[1] Politecnico di Torino, {name.surname}@polito.it

[2] H-Farm Innovation, {name.surname}@h-farm.com

Abstract

The use of Natural Language Processing techniques in the legal domain has become established for supporting attorneys and domain experts in content retrieval and decision-making. However, understanding the legal text poses relevant challenges in the recognition of domain-specific entities and the adaptation and explanation of predictive models. This paper addresses the Legal Entity Name Recognition (L-NER) and Court judgment Prediction (CPJ) and Explanation (CJPE) tasks. The L-NER solution explores the use of various transformer-based models, including an entity-aware method attending domain-specific entities. The CJPE proposed method relies on hierarchical BERT-based classifiers combined with local input attribution explainers. We propose a broad comparison of eXplainable AI methodologies along with a novel approach based on NER. For the L-NER task, the experimental results remark on the importance of domain-specific pre-training. For CJP our lightweight solution shows performance in line with existing approaches, and our NER-boosted explanations show promising CJPE results in terms of the conciseness of the prediction explanations.

1 Introduction

AI-powered tools for legal AI can analyze vast amounts of documents providing lawyers and judges with relevant insights on case laws (Sansone and Sperlí, 2022). For example, based on the analysis of past court cases legal AI models can predict the outcome of similar cases, thus offering efficient and effective ways of resolving disputes (Medvedeva et al., 2022).

The legal domain poses relevant challenges due to the specificity of the legal language, the variety of scenarios and application contexts, and the temporal evolution of norms and regulations (Chalkidis and Søgaard, 2022; Benedetto et al., 2022).

This paper addresses two notable legal AI tasks, i.e., Legal Named Entity Recognition and Court Judgment Prediction with Explanation.

Legal Named Entity Recognition The L-NER task aims at annotating portions of legal content with domain-specific entities. Annotations are particularly helpful to support content retrieval and indexing. For example, the identification of relevant legal provisions can help lawyers and judges make informed decisions based on previous cases. Despite NER being a well-known NLP task, the application of existing techniques to legal data is challenged by the inherent complexity and nuances of the language used in the legal domain (Williams, 2005; Tiersma, 2000). We address L-NER by exploring the use of established transformer-based models, established for many NLP tasks (Vaswani et al., 2017) and tailored to the legal domain (Chalkidis et al., 2020). Unlike previous studies, we also explore the use of the *entity-aware attention mechanism* (Yamada et al., 2020), which allows the transformer to also attend to domain-specific entities. The preliminary results achieved on the development set confirm the potential benefits of leveraging the entity-aware attention mechanism in L-NER.

Court Judgment Prediction with Explanation CJPE focuses on predicting the outcome of a given case. Unlike the traditional Legal judgment Prediction (Cui et al., 2022), it entails predicting the outcome along with a textual explanation consisting of an extract of the case content. The presented approach to CJPE relies on hierarchical transformer-based models fine-tuned on annotated legal judgments. On top of the classification model, a post-hoc feature attribution method is exploited to derive sentence-level importance for judgment prediction. Beyond testing multiple models and fine-tuning datasets, we empirically analyze two complementary aspects, i.e., the role of sentence-level tokenization and the use of single- or mixed-type train-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1401–1411
July 13-14, 2023

ing data and models. The experiments show that single-type models on average perform best due to the high specificity of the legal vocabulary and syntax. Furthermore, sentence-level tokenization allows transformers to effectively handle long documents, enhancing conciseness of the generated explanations.

This paper is organized as follows:

- In Section 2 we examine the related works and highlight the difference between the present work and existing approaches.
- In Section 3 we provide an overview of the methodologies employed for the three tasks.
- In Section 4 we describe the experimental setup, and the metrics, and we provide an extensive validation and discussion of the results.
- Section 5 draws the conclusion and discusses our future research lines.

2 Related Works

Transformer-based models have exhibited remarkable performance in many legal AI domains such as legal question answering (Hendrycks et al., 2021) and legal document summarization (Jain et al., 2021). This paper addresses three specific research lines in Legal AI: Named Entity Recognition, Court judgment Prediction, and Predictions Explanation.

2.1 Legal Named Entity Recognition

Prior works on Named Entity Recognition have already explored the use of statistical models (e.g., (McCallum and Li, 2003)) and, more recently, of deep neural networks (Li et al., 2020). Pioneering works on Legal NER focus on named entity recognition and resolution on US case law (Dozier et al., 2010) and on the creation of a German NER dataset with fine-grained semantic classes (Leitner et al., 2020). Since deep learning approaches require large-scale annotated datasets and general-purpose NER models are trained on a different set of entities, publicly available legal NER data sets have recently been made available (e.g., (Au et al., 2022)). Inspired by the recent advances in NER tasks with span representation (Ouchi et al., 2020), in this paper we explore the use of entity-aware attention mechanism (Yamada et al., 2020) to accomplish the L-NER task.

2.2 Court Judgment Prediction

The problem of predicting court case outcomes has received considerable attention in recent years (Cui et al., 2022). Most research efforts on the jurisprudence of the U.S. Supreme Court (Strickson and De La Iglesia, 2020; Kowsrihawat et al., 2018). Other studies analyze the cases of the European Court of Human Rights, utilizing both traditional and machine learning methods (Aletras et al., 2016; Visentin et al., 2019; Quemy and Wrembel, 2020). Still others are related to the judicial system in India (Shaikh et al., 2020; Malik et al., 2021).

Transformer-based models have recently shown to achieve remarkable results (Chalkidis et al., 2019, 2020; Kaur and Bozic, 2019; Medvedeva et al., 2021). To the best of our knowledge, none of them focus on handling long legal documents using attention-based models such as Longformer (Beltagy et al., 2020) for court judgment prediction or hierarchical-version of transformer model (Lu et al., 2021), where the hierarchical mechanism is based on the attention.

2.3 Predictions Explanation in the Legal Domain

The growing adoption of automatic decision-making systems raised awareness of its risks. Research on explainable AI has consequently grown, addressing the need to understand model behavior (Adadi and Berrada, 2018). Hence, several XAI approaches have been proposed in the literature (Lundberg and Lee; Ribeiro et al.; Pastor and Baralis; Sarti et al.; Simonyan et al.; Sundararajan et al.; Ventura et al.; Wallace et al., inter alia) and multiple approaches and analysis have been designed to assess their quality (Atanasova et al., 2020; Attanasio et al., 2022; DeYoung et al., 2020; Jacovi and Goldberg, 2020). The demand for explainability is particularly imperative in the legal domain (Bibal et al., 2021).

A line of work consists in directly adopting interpretable models. To judge the violation of an article of the convention of human rights, Aletras et al. train an SVM with a linear kernel on n-grams and topics to facilitate the model interpretation. Given their outstanding performance, black-box models as deep learning models, especially transformer-based ones, are, however, more widely adopted. As a result, multiple works leverage post-hoc explanation methods to explain the reasons behind individual predictions of black-box models. Górski

et al. adopt the explanation method Grad-CAM, firstly proposed for computer vision, to understand predictions of convolutional neural networks for legal texts. Górski and Ramakrishna compares Grad-CAM (Selvaraju et al., 2017), LIME (Ribeiro et al., 2016), and SHAP (Lundberg and Lee, 2017) explanation methods for a legal text classification problem. The work assesses the plausibility of explanations compared to the judgment of legal professionals. The work of Malik et al. also applies a post-hoc explanation method and evaluates the plausibility of the derived explanation compared to the human rationales. The authors propose an occlusion-based approach leveraging masking individual chunk embedding and estimating their importance via classification probability change. In our work, we assess the quality of three explanation methods, i.e., a sentence-occlusion-based approach, gradient, and gradient-per-input explanation methods. We then propose a novel approach to enhance explanations with Named Entity Recognition importance tagging.

3 System Overview

In this section we separately introduce the approaches to L-NER, CJP (classification only), and CJPE (classification with predictions explanation).

3.1 The L-NER Task

Problem statement The L-NER task (Kalamkar et al., 2022) aims at identifying and classifying the named entities in unstructured legal texts. Specifically, we identify the following entities: the name of the court, the name of petitioners, the name of respondents, the name of the judge, the name of the lawyers, the date of the judgment, any organization involved, geopolitical locations, name of the act or law, sections, subsections and articles, past cases, case number, name of witnesses and other person involved.

Contribution Our main contributions to L-NER is twofold:

- We apply a recently proposed *entity-aware attention mechanism*, implemented in the LUKE model (Yamada et al., 2020). It produces contextualized representations of both words and entities at the same time.
- We carry out a comprehensive analysis of the performance of transformer-based models for L-NER. In particular, we examine the effect of using both domain-specific pre-training and fine-tuned models. The purpose is to investigate the impact of pre-training on the quality of the models' representations, the effectiveness of transfer learning in the legal domain, and the ability of the pre-trained models to capture the nuances of the legal language and domain-specific vocabulary.

Beyond LUKE (Yamada et al., 2020), hereafter we will consider the established BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models generically fine-tuned for the NER task (namely, *BERT-large* and *RoBERTa-large*) or specifically on the legal domain (*Legal BERT-base* and *Legal RoBERTa-base*), EURLEX (*BERT-base*), and ECHR (*BERT-base*) corpora (Chalkidis et al., 2020). All the pre-trained checkpoints of these models are taken from the Hugging Face hub repository[1].

3.2 The CJP Task

Problem statement The task of Court Judgment Prediction (CJP) and Explanation (Malik et al., 2021) aims to predict the decision for a case given all its facts and arguments (CJP). It returns a binary outcome per case.

Contribution The proposed methodology for the CJP task consists of three main stages. Firstly, we employ four distinct transformer-based encoders, namely *RoBERTa-base*, *RoBERTa-large*, *Legal BERT*, and *Longformer-4096*, to generate alternative document embeddings. All the pre-trained checkpoints of these models are taken from the Hugging Face hub repository[1]. The two types of train documents, i.e., single and multi, have been analyzed separately. In the multi-type dataset, for each case, multiple petitions have been filed by the appellant leading to multiple decisions. As stated by the author, this dataset is a superset of the single-type dataset (where for each case a single petition have been filed).

As these models, including *Longformer*, have a limited capacity for token input compared to the average length of legal documents, we investigate which section of the document exerts the most significant influence on classification performance, either the head or the tail, as previously studied by (Malik et al., 2021).

[1] https://huggingface.co/models latest access: February 2023

The second stage relies on a hierarchical strategy, wherein we segment sentences based on document syntax before encoding them using the best performing encoder identified at the previous step. We use *ply* library and customize the sentence splitter in order to avoid split occurring in presence of abbreviations or digits. The hierarchical model comprises a stack of attention layers followed by a stack of linear layers. The resulting embeddings for each document are combined through an average pooling operation between sentence representations, and the document representation is subsequently passed to the linear layers to compute the final prediction.

Finally, we generate an ensemble prediction exclusively for the test set by combining different sets of hierarchical models, trained using distinct document types and settings. The ensemble method maximizes the per-class sum of the probabilities returned by the considered models. This approach leverages the models' diversity to enhance the final prediction's robustness and accuracy.

3.3 The CJPE Task

Problem statement The task of Court Judgment Prediction and Explanation (CJPE) (Malik et al., 2021) combines the prediction of the final decision (binary outcome) with a prediction explanation consisting of a subset of case sentences that justify the decision. To generate the predictions CJPE re-uses the same approach and derived models used for the CJP sub-task (see Section 3.2).

Contribution We leverage post-hoc input attribution explanation methods to generate predictions explanations. This class of methods can be applied to explain individual predictions of a generic model. Post-hoc feature attribution methods, given a model, a target class, and a prediction, measure how much each token contributed to that outcome. These approaches are typically adopted by providing input tokens as input, and they provide as output the contribution to the prediction of each token. However, this scenario is not suitable for our case. First, the CJPE task aims at deriving explanations at the sentence level to understand which sentences are relevant to the prediction. Hence, we should remap token-level attributions to sentence-based ones. Second, expensive computational methods such as LIME (Ribeiro et al., 2016) may struggle due to the long length of documents when applied at the token level (Malik et al., 2021).

Our CJP already works at the sentence level: the model receives as input the entire document, split into sentences. Hence, we leverage post-hoc feature attribution methods to derive the importance of the input sentences provided as input. Specifically, we adopt Gradient (Simonyan et al., 2013) (also known as Saliency), Integrated Gradient (Sundararajan et al., 2017), and leave-one-out methods. At the implementation level, we use *ferret* (Attanasio et al., 2023), a XAI library that generates and benchmarks explanations for Transformers models. We extended the Explainer APIs of these approaches from *ferret* to deal with sentence inputs rather than tokens.

The adopted post-hoc attribution methods leverage only the input document and the model to derive explanations. Hence, no external knowledge of which parts of the inputs should be relevant for *human experts* is considered. Unfortunately, data on human rationales for the reasons behind decisions are typically unavailable. We propose to leverage NER tagging to enhance explanations, providing the information on which sentence contains legal entities. The intuition is that sentences containing legal entities should be more meaningful for the predicted outcome. Specifically, we assign each sentence's proportion of legal entities as a relevance score (NER score). We then boost the explanations post-hoc attribution method with the NER scores. Specifically, let $e(s)$ be the sentence attribution score provided by an attribution method for a sentence s, let $n(s)$ be its NER score, and $\beta \in \mathbb{R}_{\geq 0}$ a boosting parameter. The boosted sentence score is derived as $e(s)(1 + \beta n(s))$. The larger the value β, the more importance is given to the presence of legal entities in the sentence.

4 Experimental Results

In this section, we report the outcomes of our empirical investigations on the three aforementioned tasks. Specifically, we participated in the shared task competition for tasks B, C-1, and C-2 (Modi et al., 2023) and evaluated the performance of our proposed models on the respective test sets using the official evaluation metrics.

We rank 11th out of 17 participants to the L-NER task, 4th out of 11 participants to CJP, and 5th out of 11 to CJPE. These findings offer valuable insights into the potential of the proposed models for enhancing the accuracy and efficiency of legal NLP applications.

Table 1: Results of Task L-NER (B). Note that *ft* indicates the fine-tuning step.

Model	Dev Set				Test Set
	F1 Strict	F1 Partial	F1 Exact	F1 Type Match	F1 score
BERT-large (ft on NER)	83.96%	89.64%	85.37%	90.95%	78.94%
RoBERTa-large (ft on NER)	88.38%	92.80%	89.63%	93.56%	74.83%
LegalBERT-base	87.76%	92.15%	88.70%	93.41%	83.09%
LegalRoBERTa-base	86.39%	91.36%	87.66%	92.47%	73.17 %
BERT-base (ft on EURLEX)	86.34%	91.66%	87.83%	92.53%	82.91%
BERT-base (ft on ECHR)	86.77%	91.76%	88.13%	92.65%	**83.18%**
LUKE-base	88.89%	92.73%	89.85%	93.49%	75.40%
LUKE-large	**89.88%**	**93.45%**	**90.68%**	**94.20%**	76.60%

Table 2: Results of Task CJP (C1). Unique hierarchical models.

Document Type	Linear Layers	Dev Set			Test Set		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Single	3	**78.10%**	**77.97%**	**77.94%**	66.84%	66.73%	66.79%
Single	5	76.74%	76.56%	76.52%	**68.06%**	66.20%	67.11%
Single	7	77.58%	77.06%	76.95%	67.80%	**67.63%**	**67.71%**
Multi	3	76.54%	76.36%	76.31%	65.49%	65.12%	65.30%
Multi	5	**78.03%**	77.26%	77.11%	65.46%	65.16%	65.31%
Multi	7	**78.03%**	**77.87%**	**77.83%**	66.86%	63.61%	65.19%

4.1 Experimental Design

Hardware Experiments were run on a machine equipped with Intel® Core™ i9-10980XE CPU, 2 × Nvidia® RTX A6000 GPU, 128 GB of RAM running Ubuntu 22.04 LTS. We provide detailed information about the models used for the evaluation and the fine-tuning procedure in the official project repository[2].

Parameter setting For the L-NER task, the *LUKE-based* models are trained using a batch size of 256 with a 10^{-4} learning rate, while the *BERT-based* models utilize a batch size of 1 and a learning rate of 10^{-4}. Both models are trained for a maximum of 5 epochs with an early stopping criterion. Additionally, a weight decay of 0.01 is applied for both models, and the warmup ratio is set to 0.06.

For the CJP task, we train sentence encoders for a maximum of 15 epochs, using a learning rate of $5 \cdot 10^{-5}$, a warmup ratio of 0.06, a weight decay of 0.01, and a batch size of 64. The hierarchical transformer-based architecture has a maximum length of 256 tokens and it is trained for a maximum of 100 epochs, with a learning rate of $5 \cdot 10^{-5}$, and a batch size of 256.

[2] https://github.com/koudounasalkis/PoliToHFI-SemEval2023-Task6

4.2 Evaluation Metrics

L-NER On the development set, the models are evaluated by using strict, partial, exact, and type-match F1 scores on the combined preamble and judgment sentences:

- *Strict*: exact boundary surface string match and entity type;
- *Exact*: Exact match of the entity's boundaries to the corresponding boundaries in the text, without considering the entity's type;
- *Partial*: Surface string match that covers only a portion of the boundary, irrespective of its type;
- *Type-match*: Some overlap between the tagged entity and the gold entity is required along with entity type match; this score gives an indication of how much overlap exists between ground truth and prediction.

On test data, we report the standard F1-score (Kalamkar et al., 2022).

CJP We evaluate the binary court judgment predictions using the macro Precision, Recall and F1 score metrics.

CJPE We quantitatively evaluate explanations by comparing them with experts' gold annotations. The gold annotations, collected by the task proposers, are the set of sentences considered relevant by legal experts. These ground-truth explanations are unavailable at training/test time and could not be used as input to the model or to tune explanation parameters. To measure the adherence of our explanations with gold ones, we use the ROUGE-L, ROUGE-1, ROUGE-2 (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), Jaccard Similarity, Overlap Maximum, and Overlap Minimum (Malik et al., 2021) metrics.

4.3 Results

L-NER Table 1 compares the performance of several models for L-NER on both the development and test sets. The empirical findings of this study suggest that models fine-tuned on legal documents outperform the ones fine-tuned on general-purpose text. The *BERT-base* model fine-tuned on the European Court of Human Rights (ECHR) dataset achieved the best performance on the test set with an F1 score of 83.18%, close to the *Legal BERT-base* outcome (F1 score: 83.09%). A possible reason is that the fine-tuned representation achieved on ECHR is robust enough to well describe other legal datasets as well. In fact, the ECHR dataset contains a large number of legal judgments, capturing a broad range of legal concepts and terms, many of them are also relevant to the Indian legal documents. Notably, the performance gaps between the best performing models are relatively small. While *LUKE-base* and *LUKE-large* achieve the best performance on the development set, they exhibit the worst performance on the test data. This could be due to models pre-training on a general-purpose corpus, which may be not always effective in accomplishing a specific task. This confirms the importance of fine-tuning on domain-specific data.

The results of this study underscore the superior efficacy of models fine-tuned on legal text data for L-NER tasks. In the presence of legal pre-training, a BERT model of the same architecture size shows a +4% improvement over its counterpart. Additionally, *Legal RoBERTa-base*, which has a smaller architecture than *RoBERTa-large*, outperform the latter in some of the F1 scores, further demonstrating the importance of pre-training on legal text data.

Table 3: Results of Task CJP (C1). Model ensembles.

Ensemble Type	Dev Set Precision	Dev Set Recall	Test Set F1 score
Top_{S+M}	**67.80%**	**67.63%**	**67.71%**
All_S	**67.80%**	**67.63%**	**67.71%**
All_M	66.86%	63.61%	65.19%
All_{S+M}	66.86%	63.61%	65.19%

CJP The preliminary results confirm the higher importance of the document tail compared to other document sections for the CJP task. The set of results obtained using various document types and textual encoders confirm the preliminary findings reported by (Malik et al., 2021).

The experimental results also indicate that the encoders' performance depends on the document type. Specifically, *RoBERTa-large* and *Legal BERT* have shown to be the most effective models for single-type and multi-type documents, respectively.

Based on the preliminary results, we employ the aforementioned encoders to compute the sentence embeddings provided as input to the hierarchical models. Table 2 summarizes the hierarchical model outcome for both single- and multi-type documents. Each hierarchical model is a series of linear layers atop two attention layers and is validated on a set of documents of the same kind as those used in the training. Notice that the test document type is unknown.

The performance of the hierarchical models in the validation phase, i.e., between 76% and 78%, are comparable to that of the best baseline. The main difference is in the higher efficiency of attention-based networks compared to BiGRUs.

It also appears that the number of linear network layers has a weak influence on CPJ performance. Such a finding is confirmed by further tests on the transformer architectures.

It is worth noting that the model performance slightly degrades on the test set, independently of document type and distribution (i.e., roughly 10% drop compared to the development set).

Table 3 shows the outcomes of the model ensembles utilized for the CJP task. Specifically, Top_{S+M} refers to the ensemble comprising the best performing single- and multi-type models, whereas All_S, All_M, and All_{S+M} are ensembles composed of all single-type, all multi-type, and all models, respectively (see Table 2). These ensembles are evaluated on the test set to ascertain their effectiveness in

Table 4: Results for CJPE (Task C). Quality of the explanation of LOO (Leave-one-out), GradientXInput (GxI), Gradient (G), and gradient β-boosted via NER (G+NER-β) considering X% of the sentences as explanation.

Explainer	ROUGE 1	ROUGE 2	ROUGE L	Jaccard	overlap min	overlap max	BLEU	METEOR
LOO 40%	0.1963	0.0430	0.1730	0.1132	0.4036	0.1458	0.0738	0.2194
GxI 40%	0.1956	0.0433	0.1727	0.1126	**0.4155**	0.1439	0.0711	0.2179
G 40%	0.2010	0.0446	0.1785	0.1168	0.4089	0.1509	0.0759	**0.2231**
G+NER-3 40%	0.2009	0.0447	0.1784	0.1167	0.4091	0.1508	0.0758	0.2228
G 30%	0.2052	0.0452	0.1816	0.1201	0.3748	0.1620	0.0865	0.2210
G+NER-5 30%	0.2042	0.0452	0.1805	0.1193	0.3766	0.1600	0.0851	0.2217
G 25%	0.2071	0.0451	0.1819	0.1219	0.3532	**0.1692**	0.0920	0.2141
G+NER-5 25%	**0.2074**	**0.0454**	**0.1821**	**0.1220**	0.3548	**0.1692**	**0.0922**	0.2145

handling the diverse document types. Notice that the model ensembles do not lead to any improvements in performance on the test set, indicating that the unknown typology of test documents cannot be effectively managed through ensembling.

CJPE We consider the best performing model on ILDC documents where multiple petitions have been filed by the appellant leading to multiple decisions (multi). As stated by the author, this dataset is a superset of the single-type dataset, therefore, we focus our analyses on models trained to predict multiple decisions for achieving higher generalizability than the single-type setup with almost comparable results.

The test set for the CJPE consists of 50 documents annotated by domain experts with gold labels and explanations (unavailable at test time). The adopted model achieves an F1-score of 45.25%.

We evaluate the quality of our leave-one-out (LOO), gradient (G), and gradientXInput (GxI) explanations and the one boosted with NER (Explainer-NER-β). We compare the explanation's quality with experts' gold annotations.

Given the input document, the adopted explainers provide an importance score for each document sentence for the predicted class. We derive the explanation as the set of top-K sentences by importance. Table 4 shows a summary of the results for the submitted explanations. Following (Malik et al., 2021), we first consider 40% of the sentences as explanations. Among LOO, GxI, and G, G explanations achieve higher results for all metrics except minimum overlap. Hence, in the following evaluations, we consider only Gradient explanations. We boosted gradient explanations with NER tagging leveraging LUKE-large model predictions. This choice has been made according to the validation results in Table 1. We set the value β for the boosting to 3. It corresponds to a modification of the relevant sentences in the explanations of 9% across all documents and 24% of the documents. In this case, we do not observe a significant impact of the NER boosting.

We then study the impact of the length of explanations. We prefer concise explanations to summarize the relevant content better and to ease understanding. We evaluate explanations that represent from 25% to 30% of the documents. We observe that shorter explanations are indeed associated with higher evaluation scores. From the empirical results we further observe that NER boosting is effective for short explanations. When selecting a few sentences, the NER tagging provides insights into the importance of sentences containing legal entities that domain experts might consider as part of the decision justification.

5 Conclusions and Future Research Directions

This paper addressed two established legal AI task, i.e., the Named Entity Recognition and Court judgment Prediction. More specifically,

- It proposed to apply LUKE to address Named Entity Recognition on legal data sources and performed a comprehensive fine-tuning of state-of-the-art language models tailored to the legal domain.
- It extensively fine-tuned state-of-the-art language models to improve their ability to predict the outcome of court cases, proposing an approach based on hierarchical, attention-based models, and a fine-grained sentence splitting.

- It presented a novel approach based on NER to explain the CJP predictions. By leveraging NER tagging, we enhanced the explanations produced by well-known XAI methodologies by also considering the presence of legal entities within a sentence.

In light of the achieved results, the L-NER models fine-tuned on legal text data outperformed those fine-tuned on general text data. Specifically, the *BERT-base* model fine-tuned on the European Court of Human Rights (ECHR) dataset performed best. The hierachical approach to CJP achieved results comparable to the baseline methods. For the CJPE task, NER-boosted explanations show promising results in producing concise yet informative explanations.

As future work, we plan to explore the applicability of entity-aware transformers to other legal AI tasks such as relation extraction. Entities could help to better capture the inter-dependencies between different entities and their relationships within the legal text.

Limitations and Ethics Statement

CJP systems intend to assist legal experts by providing useful information and not substituting them. The performance results that we and the other participants of the task obtained show that much research effort should be made to entitle this aim. CJP systems may reflect biases and discriminatory aspects of our society. The explanations behind CJP systems' predictions could help practitioners reveal, mitigate, and remove biases in such systems. Moreover, explanations allow legal experts to assess the reason behind predictions, assessing if the system decisions are for the right reasons. It is then essential that system explanations are of adequate quality. The preliminary study of the plausibility of explanations we carried out in this work goes in this direction. We envision a more comprehensive assessment of the plausibility and faithfulness of explanations (Jacovi and Goldberg, 2020). Moreover, we encourage evaluating CJP performance also at the subgroup level (Pastor et al., 2021a,b; Goel et al., 2021) to assess which data subgroups experience lower performance. Subgroup-level evaluation has proven beneficial in identifying modeling issues or biases toward specific subgroups for various transformer-based models (Koudounas et al., 2023).

Acknowledgements

This work has been partially supported by the "National Centre for HPC, Big Data and Quantum Computing", CN000013 (approved under the M42C Call for Proposals - Investment 1.4 - Notice "Centri Nazionali" - D.D. No. 3138, 16.12.2021, admitted for funding by MUR Decree No. 1031,17.06.2022), the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), and SmartData@PoliTO center on Big Data and Data Science. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Ting Wai Terence Au, Ingemar J. Cox, and Vasileios Lampos. 2022. E-NER - an annotated named entity recognition corpus of legal text. *CoRR*, abs/2212.09306.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Irene Benedetto, Luca Cagliero, and Francesco Tarasconi. 2022. Automatic inference of taxonomy relationships among legal documents. In *New Trends in Database and Information Systems*, pages 24–33, Cham. Springer International Publishing.

Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29:149–169.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

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

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named entity recognition and resolution in legal text*. Springer.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Łukasz Górski and Shashishekar Ramakrishna. 2021. Explainable artificial intelligence, lawyer's perspective. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 60–68, New York, NY, USA. Association for Computing Machinery.

Łukasz Górski, Shashishekar Ramakrishna, and Jędrzej M. Nowosielski. 2021. Towards grad-cam based explainability in a legal text processing pipeline. extended version. In *AI Approaches to the Complexity of Legal Systems XI-XII*, pages 154–168, Cham. Springer International Publishing.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. *CoRR*, abs/2103.06268.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arshdeep Kaur and Bojan Bozic. 2019. Convolutional neural network-based automatic prediction of judgments of the european court of human rights. In *Irish Conference on Artificial Intelligence and Cognitive Science*.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2023. Exploring subgroup performance in end-to-end speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism.

2018 5th Asian Conference on Defense Technology (ACDT), pages 50–55.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. *A Sentence-Level Hierarchical BERT Model for Document Classification with Limited Labelled Data*, pages 231–241.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*, pages 12–23.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2022. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 Task 6: LegalEval: Understanding Legal Texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics (ACL).

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6459, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Eliana Pastor and Elena Baralis. 2019. Explaining black box models by means of local rules. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 510–517, New York, NY, USA. Association for Computing Machinery.

Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021a. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1400–1412, New York, NY, USA. Association for Computing Machinery.

Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. 2021b. How divergent is your data? *Proc. VLDB Endow.*, 14(12):2835–2838.

Alexandre Quemy and Robert Wrembel. 2020. On integrating and classifying legal text documents. In *Database and Expert Systems Applications*, pages 385–399, Cham. Springer International Publishing.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. *ArXiv*, abs/2302.13942.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402. International Conference on Computational Intelligence and Data Science.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 3rd International Conference on Information Science and Systems*, ICISS '20, page 204–209, New York, NY, USA. Association for Computing Machinery.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Peter Tiersma. 2000. Legal language. *Bibliovault OAI Repository, the University of Chicago Press*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. 2022. Trusting deep learning natural-language models via local and global explanations. *Knowledge and Information Systems*, 64(7):1863–1907.

Andrea Visentin, Alessia Nardotto, and Barry O'Sullivan. 2019. Predicting judicial decisions: A statistically rigorous approach and a new ensemble classifier. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1820–1824.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Christopher Williams. 2005. *Tradition and Change in Legal English*. Peter Lang Verlag, Lausanne, Switzerland.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.