Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases

Ilias Chalkidis [† ‡] **Manos Fergadiotis** [† ‡] **Dimitrios Tsarapatsanis** [⋆]
Nikolaos Aletras [◇] **Ion Androutsopoulos** [‡ †] **Prodromos Malakasiotis** [† ‡]
[†] EY AI Centre of Excellence in Document Intelligence, NCSR "Demokritos"
[‡] Department of Informatics, Athens University of Economics and Business
[◇] Computer Science Department, University of Sheffield
[⋆] Law School, University of York

Abstract

Interpretability or *explainability* is an emerging research field in NLP. From a *user-centric* point of view, the goal is to build models that provide proper justification for their decisions, similar to those of humans, by requiring the models to satisfy additional constraints. To this end, we introduce a new application on legal text where, contrary to mainstream literature targeting word-level rationales, we conceive rationales as selected *paragraphs* in multi-paragraph structured court cases. We also release a new dataset comprising European Court of Human Rights cases, including annotations for paragraph-level rationales. We use this dataset to study the effect of already proposed rationale constraints, i.e., *sparsity*, *continuity*, and *comprehensiveness*, formulated as regularizers. Our findings indicate that some of these constraints are not beneficial in paragraph-level rationale extraction, while others need re-formulation to better handle the multi-label nature of the task we consider. We also introduce a new constraint, *singularity*, which further improves the quality of rationales, even compared with noisy rationale supervision. Experimental results indicate that the newly introduced task is very challenging and there is a large scope for further research.

1 Introduction

Model *interpretability* (or *explainability*) is an emerging field of research in NLP (Lipton, 2018; Jacovi and Goldberg, 2020). From a *model-centric* point of view, the main focus is to demystify a model's inner workings, for example targeting self-attention mechanisms (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), and more recently Transformer-based language models (Clark et al., 2019; Kovaleva et al., 2019; Rogers et al., 2020). From a *user-centric* point of view, the main focus is to build models that learn to provide proper justification for their decisions, similar to those of humans, (Zaidan et al., 2007; Lei et al., 2016; Chang et al., 2019; Yu et al., 2019) by requiring the models to satisfy additional constraints.

Here we follow a *user-centric* approach to *rationale extraction*, where the model learns to select a subset of the input that justifies its decision. To this end, we introduce a new application on legal text where, contrary to mainstream literature targeting word-level rationales, we conceive rationales as automatically selected paragraphs in multi-paragraph structured court cases. While previous related work targets mostly binary text classification tasks (DeYoung et al., 2020), our task is a highly skewed multi-label text classification task. Given a set of paragraphs that refer to the facts of each case (henceforth *facts*) in judgments of the European Court of Human Rights (ECtHR), the model aims to predict the *allegedly violated* articles of the European Convention of Human Rights (ECHR). We adopt a *rationalization by construction* methodology (Lei et al., 2016; Chang et al., 2019; Yu et al., 2019), where the model is regularized to satisfy additional constraints that reward the model, if its decisions are based on concise rationales it selects, as opposed to inferring explanations from the model's decisions in a post-hoc manner (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017; Murdoch et al., 2018).

Legal judgment prediction has been studied in the past for cases ruled by the European Court of Human Rights (Aletras et al., 2016; Medvedeva et al., 2018; Chalkidis et al., 2019) and for Chinese criminal court cases (Luo et al., 2017; Hu et al., 2018; Zhong et al., 2018), but there is no precedent of work investigating the justification of the models' decisions. Similarly to other domains (e.g., financial, biomedical), explainability is a key feature in the legal domain, which may potentially improve the trustworthiness of systems that abide by the principle of the *right to explanation* (Goodman and

Correspondence to: ihalk.aueb.gr

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 226–241
June 6–11, 2021. ©2021 Association for Computational Linguistics

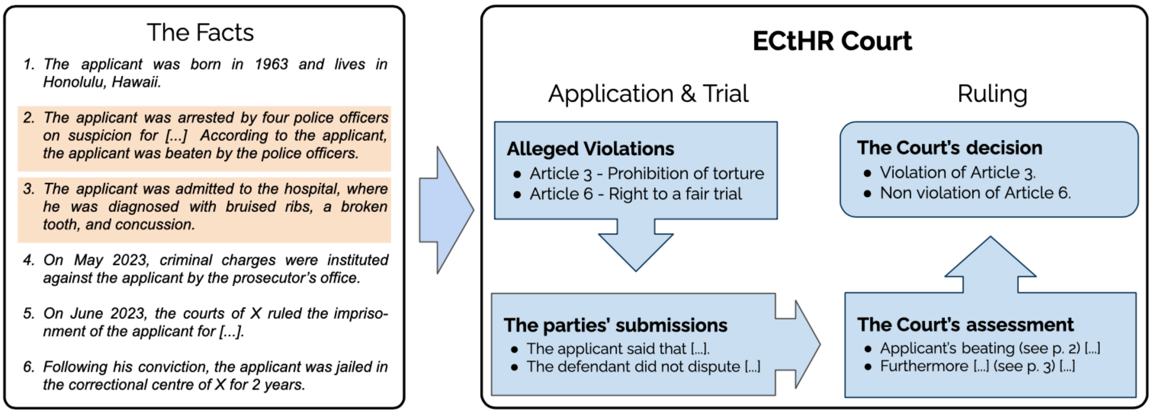

The Facts
1. The applicant was born in 1963 and lives in Honolulu, Hawaii.
2. The applicant was arrested by four police officers on suspicion for [...] According to the applicant, the applicant was beaten by the police officers.
3. The applicant was admitted to the hospital, where he was diagnosed with bruised ribs, a broken tooth, and concussion.
4. On May 2023, criminal charges were instituted against the applicant by the prosecutor's office.
5. On June 2023, the courts of X ruled the imprisonment of the applicant for [...].
6. Following his conviction, the applicant was jailed in the correctional centre of X for 2 years.
ECtHR Court
Application & Trial
Ruling
Alleged Violations
• Article 3 - Prohibition of torture
• Article 6 - Right to a fair trial
The Court's decision
• Violation of Article 3.
• Non violation of Article 6.
The parties' submissions
• The applicant said that [...].
• The defendant did not dispute [...]
The Court's assessment
• Applicant's beating (see p. 2) [...]
• Furthermore [...] (see p. 3) [...]

Figure 1: A depiction of the ECtHR process: The applicant(s) request a hearing from ECtHR regarding specific accusations (alleged violations of ECHR articles) against the defendant state(s), based on facts. The Court (judges) assesses the facts and the rest of the parties' submissions, and rules on the violation or not of the allegedly violated ECHR articles. Here, prominent facts referred in the court's assessment are highlighted.

Flaxman, 2017). We investigate the explainability of the decisions of state-of-the-art models, comparing the paragraphs they select to those of legal professionals, both litigants and lawyers, in *alleged violation prediction*. In the latter task, introduced in this paper, the goal is to predict the accusations (allegations) made by the applicants. The accusations can be usually predicted given only the facts of each case. By contrast, in the previously studied legal judgment prediction task, the goal is to predict the court's decision; this is much more difficult and vastly relies on case law (precedent cases).

Although the new task (alleged violation prediction) is simpler than legal judgment prediction, models that address it (and their rationales) can still be useful in the judicial process (Fig. 1). For example, they can help applicants (plaintiffs) identify alleged violations that are supported by the facts of a case. They can help judges identify more quickly facts that support the alleged violations, contributing towards more informed judicial decision making (Zhong et al., 2020). They can also help legal experts identify previous cases related to particular allegations, helping analyze case law (Katz, 2012). Our contributions are the following:

- We introduce *rationale extraction* for *alleged violation prediction* in ECtHR cases, a more tractable task compared to legal judgment prediction. This is a *multi-label* classification task that requires *paragraph-level* rationales, unlike previous work on word-level rationales for binary classification.
- We study the effect of previously proposed rationale constraints, i.e., *sparsity*, *continuity* (Lei et al., 2016), and *comprehensiveness* (Yu et al., 2019), formulated as regularizers. We show that *continuity* is not beneficial and requisite in paragraph-level rationale-extraction, while *comprehensiveness* needs to be re-formulated for the multi-label nature of the task we consider. We also introduce a new constraint, *singularity*, which further improves the rationales, even compared with silver (noisy) rationale supervision.
- We release a new dataset for alleged article violation prediction, comprising 11k ECtHR cases in English, with silver rationales obtained from references in court decisions, and gold rationales provided by ECHR-experienced lawyers.[1]

To the best of our knowledge, this is also the first work on rationale extraction that fine-tunes end-to-end pre-trained Transformer-based models.[2]

2 Related Work

Legal judgment prediction: Initial work on legal judgment prediction in English used linear models with features based on bags of words and topics, applying them to ECtHR cases (Aletras et al., 2016; Medvedeva et al., 2018). More recently, we experimented with neural methods (Chalkidis et al., 2019) , showing that hierarchical RNNs (Yang et al., 2016), and a hierarchical variation of BERT (Devlin et al., 2019) that encodes paragraphs, outperform linear classifiers with bag-of-word representations.

In all previous work, legal judgment prediction is tackled in an over-simplified experimental setup

[1] Our dataset is publicly available at `https://huggingface.co/datasets/ecthr_cases`, see usage example in Appendix E.

[2] Others fine-tuned such models only partially (Jain et al., 2020), i.e., top two layers, or not at all (DeYoung et al., 2020).

where only textual information from the cases themselves is considered, ignoring many other important factors that judges consider, more importantly general legal argument and past case law. Also, Aletras et al. (2016), Medvedeva et al. (2018), Chalkidis et al. (2019) treat ECtHR judgment prediction as a binary classification task per case (any article violation or not), while the ECtHR actually considers and rules on the violation of individual articles of the European Convention of Human Rights (ECHR).

In previous work (Chalkidis et al., 2019), we also attempted to predict which *particular* articles were violated, assuming, however, that the Court considers all the ECHR articles in each case, which is not true. In reality, the Court considers only alleged violations of particular articles, argued by applicants. Establishing which articles are allegedly violated is an important preliminary task when preparing an ECtHR application. Instead of oversimplifying the overall judgment prediction task, we focus on the preliminary task and use it as a test-bed for generating paragraph-level rationales in a multi-label text classification task for the first time.

Legal judgment prediction has also been studied in Chinese criminal cases (Luo et al., 2017; Hu et al., 2018; Zhong et al., 2018). Similarly to the literature on legal judgment prediction for ECtHR cases, the aforementioned approaches ignore the crucial aspect of justifying the models' predictions.

Given the gravity that legal outcomes have for individuals, explainability is essential to increase the trust of both legal professionals and laypersons on system decisions and promote the use of supportive tools (Barfield, 2020). To the best of our knowledge, our work is the first step towards this direction for the legal domain, but is also applicable in other domains (e.g., biomedical), where justifications of automated decisions are essential.

Rationale extraction by construction: Contrary to earlier work that required supervision in the form of human-annotated rationales (Zaidan et al., 2007; Zhang et al., 2016), Lei et al. (2016) introduced a *self-supervised* methodology to extract rationales (that supported aspect-based sentiment analysis predictions), i.e., gold rationale annotations were used only for evaluation. Furthermore, models were designed to produce rationales *by construction*, contrary to work studying saliency maps (generated by a model without explainability constraints) using gradients or perturbations at inference time (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017; Murdoch et al., 2018). Lei et al. (2016) aimed to produce short coherent rationales that could replace the original full texts, maintaining the model's predictive performance. The rationales were extracted by generating binary masks indicating which words should be selected; and two additional loss regularizers were introduced, which penalize long rationales and sparse masks (that would select non-consecutive words).

Yu et al. (2019) proposed another constraint to ensure that the rationales would contain all the relevant information. They formulated this constraint through a *minimax* game, where two players, one using the predicted binary mask and another using the complement of this mask, aim to correctly classify the text. If the first player fails to outperform the second, the model is penalized. Chang et al. (2019) use a Generative Adversarial Network (GAN) (Goodfellow et al., 2014), where a generator producing factual rationales competes with a generator producing counterfactual rationales to trick a discriminator. The GAN was not designed to perform classification. Given a text and a label it produces a rationale supporting (or not) the label.

Jain et al. (2020) decoupled the model's predictor from the rationale extractor to produce inherently faithful explanations, ensuring that the predictor considers only the rationales and not other parts of the text. Faithfulness refers to how accurately an explanation reflects the true reasoning of a model (Lipton, 2018; Jacovi and Goldberg, 2020).

All the aforementioned work conceives rationales as selections of words, targeting binary classification tasks even when this is inappropriate. For instance, DeYoung et al. (2020) and Jain et al. (2020) over-simplified the task of the multi-passage reading comprehension (MultiRC) dataset (Khashabi et al., 2018) turning it into a binary classification task with word-level rationales, while sentence-level rationales seem more suitable.

Responsible AI: Our work complies with the ECtHR data policy. By no means do we aim to build a 'robot' lawyer or judge, and we acknowledge the possible harmful impact (Angwin et al., 2016; Dressel and Farid, 2018) of irresponsible deployment. Instead, we aim to support fair and explainable AI-assisted judicial decision making and empirical legal studies. We consider our work as part of ongoing critical research on responsible AI (Elish et al., 2021) that aims to provide explainable and fair systems to support human experts.

	Cases	Sparsity	#Allegations
Train	9K	24%	1.8
Development	1K	30%	1.7
Test	1K	31%	1.7

Table 1: Statistics of the new ECtHR dataset. 'Sparsity' is the average percentage of paragraphs included in the silver rationales. '#Allegations' is the average number of allegedly violated articles.

3 The New ECtHR Dataset

The court (ECtHR) hears allegations regarding breaches in human rights provisions of the European Convention of Human Rights (ECHR) by European states (Fig. 1).[3] The court rules on a subset of all ECHR articles, which are predefined (alleged) by the applicants (*plaintiffs*). Our dataset comprises 11k ECtHR cases and can be viewed as an enriched version of the ECtHR dataset of Chalkidis et al. (2019), which did not provide ground truth for alleged article violations (articles discussed) and rationales. The new dataset includes the following:

Facts: Each judgment includes a list of paragraphs that represent the facts of the case, i.e., they describe the main events that are relevant to the case, in numbered paragraphs. We hereafter call these paragraphs *facts* for simplicity. Note that the facts are presented in chronological order. Not all facts have the same impact or hold crucial information with respect to alleged article violations and the court's assessment; i.e., facts may refer to information that is trivial or otherwise irrelevant to the legally crucial allegations against *defendant* states.

Allegedly violated articles: Judges rule on specific accusations (allegations) made by the applicants (Harris, 2018). In ECtHR cases, the judges discuss and rule on the violation, or not, of specific articles of the Convention. The articles to be discussed (and ruled on) are put forward (as alleged article violations) by the applicants and are included in the dataset as ground truth; we identify 40 violable articles in total.[4] In our experiments, however, the models are not aware of the allegations. They predict the Convention articles that will be discussed (the allegations) based on the case's facts, and they also produce rationales for their predictions. Models of this kind could be used by potential applicants to help them formulate future allegations (articles they could claim to have been violated), as already noted, but here we mainly use the task as a test-bed for rationale extraction.

Violated articles: The court decides which allegedly violated articles have indeed been violated. These decisions are also included in our dataset and could be used for full legal judgment prediction experiments (Chalkidis et al., 2019). However, they are not used in the experiments of this work.

Silver allegation rationales: Each decision of the ECtHR includes references to facts of the case (e.g., "*See paragraphs 2 and 4.*") and case law (e.g., "*See Draci vs. Russia (2010).*"). We identified references to each case's facts and retrieved the corresponding paragraphs using regular expressions. These are included in the dataset as silver allegation rationales, on the grounds that the judges refer to these paragraphs when ruling on the allegations.

Gold allegation rationales: A legal expert with experience in ECtHR cases annotated a subset of 50 test cases to identify the relevant facts (paragraphs) of the case that support the allegations (alleged article violations). In other words, each identified fact justifies (hints) one or more alleged violations.[5]

Task definition: In this work, we investigate *alleged violation prediction*, a multi-label text classification task where, given the facts of a ECtHR case, a model predicts which of the 40 violable ECHR articles were allegedly violated according to the applicant(s).[4] The model also needs to identify the facts that most prominently support its decision.

4 Methods

We first describe a baseline model that we use as our starting point. It adopts the framework proposed by Lei et al. (2016), which generates rationales by construction: a *text encoder* sub-network reads the text; a *rationale extraction* sub-network produces a binary mask indicating the most important words of the text; and a *prediction* sub-network classifies a hard-masked version of the text. We then discuss additional constraints that have been proposed to improve word-level rationales, which can be added to the baseline as regularizers. We argue that one of them is not beneficial for paragraph-level rationales. We also consider variants of previous constraints that better suit multi-label classification tasks and introduce a new one.

[3] The Convention is available at https://www.echr.coe.int/Documents/Convention_ENG.pdf.

[4] The rest of the articles are procedural, i.e., the number of judges, criteria for office, election of judges, etc.

[5] For details on the annotation process and examples of annotated ECtHR cases, see Appendices C, F.

4.1 Baseline Model

Our baseline is a hierarchical variation of BERT (Devlin et al., 2019) with hard attention, dubbed HIERBERT-HA.[6] Each case (document) D is viewed as a list of facts (paragraphs) $D = [P_1, \dots, P_N]$. Each paragraph is a list of tokens $P_i = [w_1, \dots, w_{L_i}]$. We first pass each paragraph independently through a shared BERT encoder (Fig. 2) to extract context-unaware paragraph representations $P_i^{[\mathrm{CLS}]}$, using the [CLS] embedding of BERT. Then, a shallow encoder with two Transformer layers (Vaswani et al., 2017) produces contextualized paragraph embeddings, which are in turn projected to two separate spaces by two different fully-connected layers, K and Q, with SELU activations (Klambauer et al., 2017). K produces the paragraph encoding P_i^K, to be used for classification; and Q produces the paragraph encoding P_i^Q, to be used for rationale extraction. The rationale extraction sub-network passes each P_i^Q encoding independently through a fully-connected layer with a sigmoid activation to produce soft attention scores $a_i \in [0, 1]$. The attention scores are then binarized using a 0.5 threshold, leading to hard attention scores z_i ($z_i = 1$ iff $a_i > 0.5$). The hard-masked document representation D_M is obtained by hard-masking paragraphs and max-pooling:

$$D_M = \mathrm{maxpool}\big([z_1 \cdot P_1^K, \dots, z_N \cdot P_N^K]\big)$$

D_M is then fed to a dense layer with sigmoid activations, which produces a probability estimate per label, $\widehat{Y} = [\hat{y_1}, \dots, \hat{y_{|A|}}]$, in our case per article of the Convention, where $|A|$ is the size of the label set. For comparison, we also experiment with a model that masks no facts, dubbed HIERBERT-ALL.

The thresholding that produces the hard (binary) masks z_i is not differentiable. To address this problem, Lei et al. (2016) used reinforcement learning (Williams, 1992), while Bastings et al. (2019) proposed a differentiable mechanism relying on the reparameterization trick (Louizos and Welling, 2017). We follow a simpler trick, originally proposed by Chang et al. (2019), where during backpropagation the thresholding is detached from the computation graph, allowing the gradients to bypass the thresholding and reach directly the soft attentions a_i.

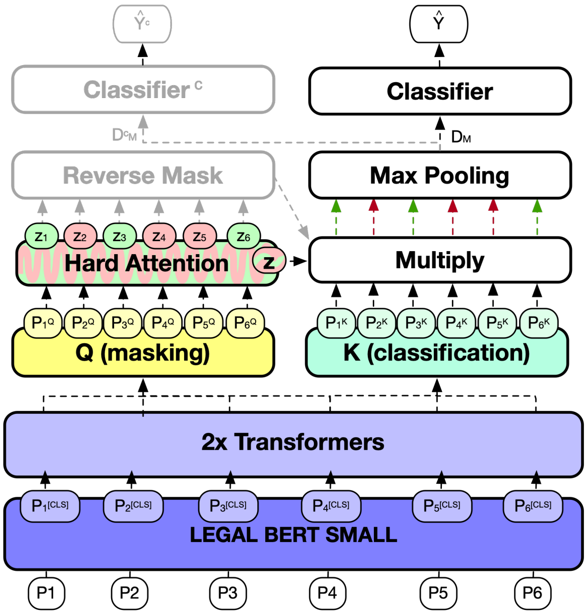

$\hat{Y}^c$
$\hat{Y}$
Classifier ^{c}
Classifier
D^c_M
D_M
Reverse Mask
Max Pooling
z_1 z_2 z_3 z_4 z_5 z_6
Hard Attention
z
Multiply
P_1^Q P_2^Q P_3^Q P_4^Q P_5^Q P_6^Q
Q (masking)
P_1^K P_2^K P_3^K P_4^K P_5^K P_6^K
K (classification)
2x Transformers
$P_1^{[CLS]}$ $P_2^{[CLS]}$ $P_3^{[CLS]}$ $P_4^{[CLS]}$ $P_5^{[CLS]}$ $P_6^{[CLS]}$
LEGAL BERT SMALL
P1 P2 P3 P4 P5 P6

Figure 2: Illustration of HIERBERT-HA. The shaded parts operate only when Lg or Lr are used.

4.2 Rationale Constraints as Regularizers

Sparsity: Modifying the word-level sparsity constraint of Lei et al. (2016) for our paragraph-level rationales, we also hypothesize that good rationales include a small number of facts (paragraphs) that sufficiently justify the allegations; the other facts are trivial or secondary. For instance, an introductory fact like "*The applicant was born in 1984 and lives in Switzerland.*" does not support any allegation, while a fact like "*The applicant contended that he had been beaten by police officers immediately after his arrest and later during police questioning.*" suggests a violation of Article 3 "Prohibition of Torture". Hence, we use a *sparsity* loss to control the number of selected facts:

$$L_s = \left| T - \frac{1}{N} \sum_{i=1}^{N} z_i \right| \tag{1}$$

where T is a predefined threshold specifying the desired percentage of selected facts per case. We can estimate T from silver rationales (Table 1).

Continuity: In their work on word-level rationales, Lei et al. (2016) also required the selected words to be *contiguous*, to obtain more coherent rationales. In other words, the transitions between selected ($z_i = 1$) and not selected ($z_i = 0$) words in the hard mask should be minimized. This is achieved by adding the following *continuity* loss:

$$L_c = \frac{1}{N-1} \sum_{i=2}^{N} |z_i - z_{i-1}| \tag{2}$$

[6] In previous work, we proposed a hierarchical variation of BERT with self-attention (Chalkidis et al., 2019). In parallel work, Yang et al. (2020) proposed a similar Transformer-based Hierarchical Encoder (SMITH) for long document matching.

In paragraph-level rationale extraction, where entire paragraphs are masked, the continuity loss forces the model to select contiguous paragraphs. In ECtHR cases, however, the facts are self-contained and internally coherent paragraphs (or single sentences). Hence, we hypothesize that the *continuity* loss is not beneficial in our case. Nonetheless, we empirically investigate its effect.

Comprehensiveness: We also adapt the *comprehensiveness* loss of Yu et al. (2019), which was introduced to force the hard mask $Z = [z_1, \ldots, z_N]$ to (ideally) keep *all* the words (in our case, paragraphs about facts) of the document D that support the correct decision Y. In our task, $Y = [y_1, \ldots, y_{|A|}]$ is a binary vector indicating the Convention articles the court discussed (gold allegations) in the case of D. Intuitively, the complement Z^c of Z, i.e., the hard mask that selects the words (in our case, facts) that Z does not select, should not select sufficient information to predict Y. Given D, let D_M, D_M^c be the representations of D obtained with Z, Z^c, respectively; let $\widehat{Y}, \widehat{Y}^c$ be the corresponding probability estimates; let L_p, L_p^c be the classification loss, typically total binary cross-entropy, measuring how far $\widehat{Y}, \widehat{Y}^c$ are from Y. In its original form, the comprehensiveness loss requires L_p^c to exceed L_p by a margin h.

$$L_g = \max(L_p - L_p^c + h, 0) \qquad (3)$$

While this formulation may be adequate in binary classification tasks, in multi-label classification it is very hard to pre-select a reasonable margin, given that cross-entropy is unbounded, that the distribution of true labels (articles discussed) is highly skewed, and that some labels are easier to predict than others. To make the selection of h more intuitive, we propose a reformulation of L_g that operates on class probabilities rather than classification losses. The right-hand side of Eq. 3 becomes:

$$\frac{1}{|A|}\sum_{i=1}^{|A|} y_i(\hat{y}_i{}^c - \hat{y}_i + h) + (1 - y_i)(\hat{y}_i - \hat{y}_i{}^c + h) \qquad (4)$$

The margin h is now easier to grasp and tune. It encourages the same gap between the probabilities predicted with Z and Z^c across all labels (articles).

We also experiment with a third variant of comprehensiveness, which does not compare the probabilities we obtain with Z and Z^c, comparing instead the two latent document representations:

$$L_g = |\cos(D_M, D_M^c)| \qquad (5)$$

where cos denotes cosine similarity. This variant forces D_M and D_M^c to be as dissimilar as possible, without requiring a preset margin.

Singularity: A limitation of the comprehensiveness loss (any variant) is that it only requires the mask Z to be better than its complement Z^c. This does not guarantee that Z is better than *every* other mask. Consider a case where the gold rationale identifies three articles and Z selects only two of them. The model may produce better predictions with Z than with Z^c, and D_M may be very different than D_M^c in Eq. 5, but Z is still not the best mask. To address this limitation, we introduce the *singularity* loss L_r, which requires Z to be better than a mask Z^r, randomly generated per training instance and epoch, that selects as many facts as the sparsity threshold T allows:

$$\begin{aligned} L_r &= \gamma \cdot L_g(Z, Z^r) \\ \gamma &= 1 - \cos(Z^r, Z) \end{aligned} \qquad (6)$$

Here $L_g(Z, Z^r)$ is any variant of L_g, but now using Z^r instead of Z^c; and γ regulates the effect of $L_g(Z, Z^r)$ by considering the cosine distance between Z^r and Z. The more Z and Z^r overlap, the less we care if Z performs better than Z^r.

The total loss of our model is computed as follows. Again L_p is the classification loss; L_p^c, L_p^r are the classification losses when using Z^c, Z^r, respectively; and all λs are tunable hyper-parameters.

$$\begin{aligned} L = {} & L_p + \lambda_s \cdot L_s + \lambda_c \cdot L_c \\ & + \lambda_g \left(L_g + L_p^c\right) + \lambda_r \left(L_r + L_p^r\right) \end{aligned} \qquad (7)$$

We include L_p^c in Eq. 7, because otherwise the network would have no incentive to make D_M^c and $\widehat{Y}^c$ competitive in prediction; and similarly for L_p^r.

Rationales supervision: For completeness we also experimented with a variant that utilizes silver rationales for noisy rationale supervision (Zaidan et al., 2007). In this case the total loss becomes:

$$L = L_p + \lambda_{ns} \cdot \text{MAE}(Z, Z^s) \qquad (8)$$

where MAE is the mean absolute error between the predicted mask, Z, and the silver mask, Z_s, and λ_{ns} weighs the effect of MAE in the total loss.

5 Experimental Setup

For all methods, we conducted grid-search to tune the hyper-parameters λ_*. We used the Adam optimizer (Kingma and Ba, 2015) across all experiments with a fixed learning rate of 2e-5.[7] All methods rely on LEGAL-BERT-SMALL (Chalkidis et al., 2020), a variant of BERT (Devlin et al., 2019), with 6 layers, 512 hidden units and 8 attention heads, pre-trained on legal corpora. Based on this model, we were able to use up to 50 paragraphs of 256 words each in a single 32GB GPU. In preliminary experiments, we found that the proposed model relying on a shared paragraph encoder, i.e., one that passes the same context-aware paragraph representations $P_i^{[CLS]}$ to both the Q and K sub-networks, as in Fig. 2, has comparable performance and better rationale quality, compared to a model with two independent paragraph encoders, as the one used in the literature (Lei et al., 2016; Yu et al., 2019; Jain et al., 2020).[8] For all experiments, we report the average and standard deviation across five runs.

We evaluate: (a) classification performance, (b) *faithfulness* (Section 2), and (c) *rationale quality*, while respecting a given sparsity threshold (T).

Classification performance: Given the label skewness, we evaluate classification performance using *micro-F1*, i.e., for each Convention article, we compute its F1, and micro-average over articles.

Faithfulness: Recall that *faithfulness* refers to how accurately an explanation reflects the true reasoning of a model. To measure faithfulness, we report *sufficiency* and *comprehensiveness* (DeYoung et al., 2020). Sufficiency measures the difference between the predicted probabilities for the gold (positive) labels when the model is fed with the whole text ($\widehat{Y_+}^f$) and when the model is fed only with the predicted rationales ($\widehat{Y_+}$). Comprehensiveness (not to be confused with the homonymous loss of Eq. 3–5) measures the difference between the predicted probabilities for the gold (positive) labels obtained when the model is fed with the full text ($\widehat{Y_+}^f$) and when it is fed with the complement of the predicted rationales ($\widehat{Y_+}^c$). We also compare classification performance (again using *micro-F1*) in both cases, i.e., when considering *masked inputs* (using Z) and *complementary inputs* (using Z^c).

Rationale quality: Faithful explanations (of system reasoning) are not always appropriate for users (Jacovi and Goldberg, 2020), thus we also evaluate rationale quality from a user perspective. The latter can be performed in two ways. *Objective* evaluation compares predicted rationales with gold annotations, typically via *Recall, Precision, F1* (comparing system-selected to human-selected facts in our case). In *subjective* evaluation, human annotators review the extracted rationales. We opt for an objective evaluation, mainly due to lack of resources. As rationale sparsity (number of selected paragraphs) differs across methods, which affects *Recall, Precision, F1*, we evaluate rationale quality with *mean R-Precision* (mRP) (Manning et al., 2009). That is, for each case, the model ranks the paragraphs it selects by decreasing confidence, and we compute Precision@k, where k is the number of paragraphs in the gold rationale; we then average over test cases. For completeness, we also report F1 (comparing predicted and gold rationale paragraphs), although it is less fair, because of the different sparsity of different methods, as noted.

ECHR article	Training cases	Classification F1 ↑
2 - Right to life	623	78.3 ± 2.3
3 - Prohibition of torture	1740	85.9 ± 0.9
5 - Right to liberty and security	1623	81.1 ± 1.5
6 - Right to a fair trial	5437	80.1 ± 1.0
8 - Right to respect for private life	1056	72.5 ± 1.8
10 - Freedom of expression	441	77.4 ± 1.6
11 - Freedom of assembly	162	72.1 ± 3.3
13 - Right to an effective remedy	1665	29.2 ± 3.3
14 - Prohibition of discrimination	444	44.8 ± 7.4
34 - Individual applications	547	10.0 ± 5.0
46 - Binding force of judgments	187	2.6 ± 3.2
P1-1 - Protection of property	1558	77.9 ± 1.3
Rest of the articles	< 100	< 50.0
Overall performance (micro-F1)		72.7 ± 1.2

Table 2: Classification performance of HIERBERT-ALL (no mask) across ECHR articles on development data, with respect to the number of training cases (instances).

6 Experimental Results

6.1 Initial Classification Performance

Table 2 reports the classification performance of HIERBERT-ALL (no masking, no rationales), across ECHR articles. F1 is 72.5% or greater for most of the articles with 1,000 or more training instances. The scores are higher for articles 2, 3, 5, 6, because (according to the legal expert who provided the gold allegation rationales), (i) there is a sufficient number of cases regarding these articles, and (ii) the interpretation and application of these articles is more fact-dependent than those of other articles, such as articles 10 or 11 (Harris, 2018). On the other hand, although there is a fair amount of training instances for articles 13, 14, 34, and 46, these articles are triggered in a variety of ways, many of which turn on legal procedural technicalities.

[7] In preliminary experiments, we tuned the baseline model on development data as a stand-alone classifier and found that the optimal learning rate was 2e-5, searching in the set {2e-5, 3e-5, 4e-5, 5e-5}. The optimal drop-out rate was 0.

[8] See Appendix B for additional details and results.

Method	sparsity	Entire Input	Masked Input (Z)		Compl. Input (Z^c)	
	(aim: 30%)	micro-F1 ↑	micro-F1 ↑	Suff. ↓	micro-F1 ↓	Comp. ↑
RANDOM CLASSIFIER	-	30.8 ± 0.3	-			
HIERBERT-ALL (no masking)	-	**73.7 ± 0.6**	-			
HIERBERT-HA + L_s (Eq. 1) (Lei et al., 2016)	31.7 ± 1.1	73.1 ± 0.6	69.5 ± 2.4	0.063	58.8 ± 1.5	0.181
HIERBERT-HA + L_s + L_g (Eq. 3) (Yu et al., 2019)	31.4 ± 1.9	72.8 ± 0.6	68.1 ± 4.4	0.069	59.0 ± 1.5	0.171
HIERBERT-HA + L_s + L_g (Eq. 5) (ours)	31.4 ± 1.3	72.6 ± 1.5	69.8 ± 0.8	0.043	59.6 ± 2.7	0.156
HIERBERT-HA + L_s + L_r (Eq. 4, 6) (ours)	31.5 ± 0.8	72.8 ± 0.5	**70.5 ± 0.8**	**0.040**	**55.9 ± 2.8**	**0.204**
HIERBERT-HA + rationale supervizion (Eq. 8)	33.1 ± 6.0	73.1 ± 0.5	69.2 ± 1.1	0.053	56.7 ± 6.6	0.191

Table 3: *Classification performance* (classification micro-F1) and *faithfulness* results on test data. Faithfulness is measured by considering *Sufficiency* (Suff.) and *Comprehensiveness* (Comp.), i.e., how close the label probabilities of the model are when using the rationales (masked input) or the complements of the rationales (complementary input), respectively, as opposed to using the entire input. Lower Suff. (↓) and higher Comp. (↑) are better. We also report micro-F1 for the masked and complementary input; higher and lower F1, respectively, are better.

6.2 Tuning the λ Hyper-parameters

Instead of tuning simultaneously all the λ_* hyper-parameters of Eq. 7, we adopt a greedy, but more intuitive strategy: we tune one λ at a time, fix its value, and proceed to the next; λs that have not been tuned are set to zero, i.e., the corresponding regularizer is not used yet. We begin by tuning λ_s, aiming to achieve a desirable level of sparsity without harming classification performance. We set the sparsity threshold of L_s (Eq. 1) to $T = 0.3$ (select approx. 30% of the facts), which is the average sparsity of the silver rationales (Table 1). We found $\lambda_s = 0.1$ achieves the best overall results on development data, thus we use this value for the rest of the experiments.[9] To check our hypothesis that continuity (L_c) is not beneficial in our task, we tuned λ_c on development data, confirming that the best overall results are obtained for $\lambda_c = 0$.[9] Thus we omit L_c in the rest of the experiments.

6.3 Comprehensiveness/Singularity Variants

Next, we tuned and compared the variants of the comprehensiveness loss L_g (Table 4). Targeting the label probabilities (Eq. 4) instead of the losses (Eq. 3) leads to lower rationale quality. Targeting the document representations (Eq. 5) has the best rationale quality results, retaining (as with all versions of L_g) the original classification performance (micro-F1) of Table 2. Hence, we keep the L_g variant of Eq. 5 in the remaining experiments of this section, with the corresponding λ_g value (1e-3).

L_g	classification	sparsity	rationale quality	
variant	micro-F1 ↑	(aim: 30%)	F1 ↑	mRP ↑
Eq. 3	73.0 ± 0.5	31.4 ± 1.9	35.4 ± 5.8	38.4 ± 5.9
Eq. 4	**73.1 ± 0.7**	31.9 ± 1.4	30.3 ± 3.0	32.6 ± 2.6
Eq. 5	72.8 ± 0.8	31.8 ± 1.3	**38.3 ± 2.3**	**41.2 ± 2.1**

Table 4: Development results for variants of L_g (*comprehensiveness*) and varying λ_g values (omitted).

L_r	classification	sparsity	rationale quality	
variant	micro-F1 ↑	(aim: 30%)	F1 ↑	mRP ↑
Eq. 3, 6	**73.4 ± 0.8**	32.8 ± 2.8	36.9 ± 3.6	39.0 ± 3.9
Eq. 4, 6	72.5 ± 0.7	32.0 ± 1.0	**39.7 ± 3.1**	**42.6 ± 3.8**
Eq. 5, 6	72.8 ± 0.3	31.5 ± 0.9	33.0 ± 2.7	35.5 ± 2.6

Table 5: Development results for variants of L_c (*singularity*) and varying λ_r values (omitted).

Concerning the singularity loss L_r (Table 5), targeting the label probabilities (Eq. 4, 6) provides the best rationale quality, comparing to all the methods considered. Interestingly Eq. 5, which performed best in L_g (Table 4), does not perform well in L_r, which uses L_g (Eq. 6). We suspect that in L_r, where we use a random mask Z^r that may overlap with Z, requiring the two document representations D_M, D_M^r to be dissimilar (when using Eq. 5, 6) may be a harsh regularizer with negative effects.

6.4 Task Performance and Faithfulness

Table 3 presents results on test data. The models that use the hard attention mechanism and are regularized to extract rationales under certain constraints (HIERBERT-HA + L_*) have comparable classification performance to HIERBERT-ALL. Furthermore, although paragraph embeddings are contextualized and probably have some information leak for all methods, our proposed extensions in rationale constraints better approximate faithfulness, while also respecting sparsity. Our proposed extensions lead to low sufficiency (lower is better, ↓), i.e., there is only a slight deterioration in label probabilities when we use the predicted rationale instead of the whole input. They also lead to high comprehensiveness (higher is better, ↑); we see a 20% deterioration in label probabilities when using the complement of the rationale instead of the whole input. Interestingly, our variant with the singularity loss (Eq. 4, 6) is more faithful than the model that uses supervision on silver rationales (Eq. 8).

[9] Consult Appendix D for more detailed results.

Method	Silver rationales (31%)		Gold rationales (36%)	
	mRP ↑	F1 ↑	mRP ↑	F1 ↑
RANDOM RATIONALE	30.2 ± 1.1	27.8 ± 1.1	35.1 ± 1.7	30.2 ± 2.2
HIERBERT-HA				
+ L_s (Eq. 1)	43.1 ± 6.5	37.3 ± 5.4	51.9 ± 5.7	45.7 ± 5.4
+ L_s + L_g (Eq. 3)	41.0 ± 5.1	37.5 ± 6.7	48.9 ± 6.5	44.5 ± 6.8
+ L_s + L_g (Eq. 5)	43.0 ± 1.5	38.5 ± 1.9	50.9 ± 3.2	45.8 ± 3.3
+ L_s + L_r (Eq. 4, 6)	**45.1 ± 2.1**	**40.9 ± 2.5**	**53.6 ± 2.3**	**48.3 ± 1.2**
+ supervision (Eq. 8)	43.1 ± 5.0	39.1 ± 7.1	51.4 ± 6.7	46.8 ± 0.5

Table 6: *Rationale quality* results on the 50 test cases that have both silver and gold allegation rationales. Average silver/gold rationale sparsity (%) in brackets.

6.5 Rationale Quality

We now consider rationale quality, focusing on HIERBERT-HA variants without rationale supervision. Similarly to our findings on development data (Tables 4, 5), we observe (Table 6) that using (a) our version of *comprehensiveness* loss (Eq. 5) or (b) our *singularity* loss (Eq. 4, 6) achieves better results compared to former methods, and (b) has the best results. The *singularity* loss is better in both settings (silver or gold test rationales), even compared to a model that uses supervision on silver rationales. The random masking of the singularity loss, which guides the model to learn to extract masks that perform better than *any* other mask, proved to be particularly beneficial in rationale quality. Similar observations are derived given the results on the full test set considering silver rationales.[10] In general, however, we observe that the rationales extracted by all models are far from human rationals, as indicated by the poor results (mRP, F1) on both silver and gold rationales. Hence, there is ample scope for further research.

6.6 Qualitative Analysis

Quality of silver rationales: Comparing silver rationales with gold ones, annotated by the legal expert, we find that silver rationales are not complete, i.e., they are usually fewer than the gold ones. They also include additional facts that have not been annotated by the expert. According the expert, these facts do not support allegations, but are included for technical reasons (e.g., *"The national court did not accept the applicant's allegations."*). Nonetheless, ranking methods by their rationale quality measured on silver rationales produces the same ranking as when measuring on gold rationales in the common subset of cases (Table 6). Hence, it may be possible to use silver rationales, which are available for the full dataset, to rank systems participating in ECtHR rationale generation challenges.

Model bias: Low mRP with respect to gold rationales means that the models rely partially on non causal reasoning, i.e., they select secondary facts that do not justify allegations according to the legal expert. In other words, the models are sensitive to specific language, e.g., they misuse (are easily fooled by) references to health issues and medical examinations as support for Article 3 alleged violations, or references to appeals in higher courts as support for Article 5, even when there is no concrete evidence.[11] Manually inspecting the predicted rationales, we did not identify bias on demographics. Although such spurious features may be buried in the contextualized paragraph encodings ($P_i^{[\text{CLS}]}$). In general, *de-biasing* models could benefit rationale extraction and we aim to investigate this direction in future work (Huang et al., 2020).

Plausibility: *Plausibility* refers to how convincing the interpretation is to humans (Jacovi and Goldberg, 2020). While the legal expert annotated all relevant facts with respect to allegations, according to his manual review, allegations can also be justified by sub-selections (parts) of rationales. Thus, although a method may fail to extract all the available rationales, the provided (incomplete) set of rationales may still be a convincing explanation. To properly estimate plausibility across methods, one has to perform a subjective human evaluation which we did not conduct due to lack of resources.

7 Conclusions and Future work

We introduced a new application of rationale extraction in a new legal text classification task concerning alleged violations on ECtHR cases. We also released a dataset for this task to foster further research. Moreover, we compared various rationale constraints in the form of regularizers and introduced a new one (*singularity*) improving faithfulness and rationale quality in a paragraph-level setup comparing both with silver and gold rationales.

In the future, we plan to investigate more constraints that may better fit paragraph-level rationale extraction and explore techniques to de-bias models and improve rationale quality. Paragraph-level rationale extraction can be also conceived as self-supervised extractive summarization to denoise long documents, a direction we plan to explore in the challenging task of case law retrieval (Locke and Zuccon, 2018).

[10] See Appendix D for rationale quality evaluation on the full test set.

[11] See Appendix F for examples of ECtHR cases.

Acknowledgments

We would like to thank the anonymous reviewers (esp. reviewer #2) for their constructive detailed comments. Nikolaos Aletras is supported by EPSRC grant EP/V055712/1, part of the European Commission CHIST-ERA programme, call 2019 XAI: Explainable Machine Learning-based Artificial Intelligence.

References

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.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Woodrow Barfield. 2020. *The Cambridge Handbook of the Law of Algorithms*. Cambridge Law Handbooks. Cambridge University Press.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 78–87.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *Finding of EMNLP*.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019. A game theoretic approach to classwise selective rationalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).*

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.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(10).

Madeleine Clare Elish, William Isaac, and Richard Zemel, editors. 2021. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT 2021)*. Association for Computing Machinery, Online.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, page 2672–2680, Montreal, Canada.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

David John Harris. 2018. *Harris, O'Boyle, and Warbrick : Law of the European Convention on Human Rights*, 4rd edition. edition. Oxford University Press.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 487–498.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

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.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Daniel Martin Katz. 2012. Quantitative legal prediction-or-how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory Law Journal*, 62:909.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Diederik P. Kingma and Jim Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In *31th Int. Conf. on Neural Information Processing Systems*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1261–1264, New York, NY, USA. Association for Computing Machinery.

Christos Louizos and Max Welling. 2017. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the International Conference on Machine Learning (ICML 2017)*, Sydney, Australia.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Conference on Empirical Methods in Natural Language Processing (EMNLP))*, pages 2727–2736.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

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 (NeurIPS)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, page 1725–1734, New York, NY, USA. Association for Computing Machinery.

Zichao Yang et al. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1480–1489.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Haoxi Zhong, Guo Zhipeng, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549.

A Why is the new dataset any different?

The new dataset is noteworthy for three reasons:

Legal rationale dataset: This is the first rationale extraction dataset for the legal domain, where justifying decisions is essential and often requires complicated reasoning. Thus the dataset is a challenging test-bed that will boost rationale extraction research. Predicting and justifying alleged article violations is also helpful in practice and would assist legal judgment prediction (Aletras et al., 2016).

Paragraph-level rationales: Each case description is a carefully planned document of multiple paragraphs enumerating facts chronologically. Each paragraph concisely provides information at a granularity considered appropriate for legal reasoning. Accordingly, rationales must be extracted at this granularity, either selecting an entire factual paragraph or not, as opposed to most rationale extraction datasets where particular words or phrases can be selected (e.g., from product reviews).

B Baseline Model

In preliminary experiments, we found that the proposed model (41.9M parameters) (Table 7, third row) relying on a shared paragraph encoder (HIERBERT) to produce both context-aware representations and rationales, has comparable classification performance and better rationale quality compared to: (i) a model with two independent paragraph encoders (82.8M parameters) (Table 7, first row), similar to the one used in the literature (Lei et al., 2016; Yu et al., 2019; Jain et al., 2020); (ii) a model that omits the Q and K projection layers (41.4M parameters) (Table 7, second row). Recall that Lei et al. (2016), Yu et al. (2019), Jain et al. (2020) extract rationales at the word-level, and their encoders, either BILSTMs (Lei et al., 2016; Yu et al., 2019) or BERT (Jain et al., 2020), operate on that level of granularity.

Method	classification micro-F1	Silver rationales mRP	Silver rationales F1
2x HIERBERT	73.4 ± 0.6	35.1 ± 7.9	29.3 ± 8.7
1x HIERBERT excl. (Q,K)	**73.5** ± **0.7**	29.2 ± 7.9	26.4 ± 7.9
1x HIERBERT + (Q,K)	73.2 ± 0.5	**35.9** ± **4.7**	**39.0** ± **4.9**

Table 7: Results on classification performance and rationale quality on development data.

C Annotation of Gold Rationales

The full dataset has the following characteristics:

- There are 1,000 cases in the test set. These are the most recent and have been ruled from October 5, 2017 until July 7, 2019.
- The average number of facts (paragraphs) per case is 25.2 ranging from 5 to 259.
- Almost half of the cases concern applications against 6 European states (defendants): Russia (229), Turkey (122), Ukraine (80), Romania (47), Moldova (44), Lithuania (43). Number of test cases in brackets.
- The allegations in the vast majority of cases (approx. 88%) concern nine articles: '6' (394), '3' (233), '5' (197), '8' (188), 'P1-1' (155), '13' (123), '35' (107), '10' (106), '2' (76). Number of test cases in brackets.

Based on the above statistics and the opinion of the legal expert, for the gold rationales we considered a subset of 50 cases with the following characteristics:

- Each case should consist of 25 ± 10 facts.

- The cases should be as representative as possible with respect to the defendants (European states).
- The cases should have allegations in a subset of the following articles {2, 3, 5, 6}, whose interpretation is more fact-dependent based on the literature (Harris, 2018) and our presented empirical results (Table 2).

The annotation guidelines briefly were:

- The annotator (legal expert) inspects (reads) all the factual paragraphs of the case and selects one or more articles in the predefined set {2, 3, 5, 6}, that should have been argued by the applicants according to the text.
- The annotator selects the factual paragraphs that "clearly" indicate allegations for the selected article(s), annotated in the first step.

The legal expert performance compared to the gold allegedly violated articles, is 92.3% micro-F1 (Table 8). In few cases, the legal expert selected more articles (hypothesized allegations for articles 3 and 5) compared to the gold ones. As he suggested, it is a common trend for the applicants, based on the legal opinion of their attorneys, to raise allegations only for a few articles that they believe can be justified and proved to be violated, i.e, if a citizen has no concrete evidence (documents) for his torture, his lawyer may suggest him to not raise this issue in his application. The legal expert also missed a few allegations for articles 2 and 6. The best of our models, (HIERBERT-SA + L_s + L_r) achieves 87.6% micro-F1 in the same subset.

ECHR article	Expert F1	Model F1
2 - Right to life	88.9	82.4
3 - Prohibition of torture	95.5	92.7
5 - Right to liberty and security	85.7	88.0
6 - Right to a fair trial	95.0	84.2
micro-F1	92.3	87.6
macro-F1	91.3	86.8

Table 8: Classification performance of the legal expert and our best method on the 50 annotated test cases, considering only the facts of each case.

D Additional Experimental Results

For completeness, we report results on development data for *sparsity* loss (L_s) in Table 9 and *continuity* loss (L_c) in Table 10 for different values of λ_* hyper-parameters.

	classification	sparsity	rationale quality	
λ_s	micro-F1 ↑	(aim: 30%)	F1 ↑	mRP ↑
0	**73.3 ± 0.9**	90.3 ± 19.3	32.4 ± 4.0	35.1 ± 5.5
0.01	72.9 ± 0.4	30.3 ± 2.7	36.6 ± 10.8	36.8 ± 8.8
0.1	73.2 ± 0.5	31.7 ± 1.1	**39.1 ± 4.7**	**39.0 ± 4.9**
0.5	71.3 ± 0.9	37.6 ± 7.5	29.1 ± 8.2	30.0 ± 8.4
1.0	71.1 ± 1.7	35.7 ± 5.6	36.2 ± 8.1	38.7 ± 7.6

Table 9: Development results varying λ_s (*sparsity*).

	classification	sparsity	rationale quality	
λ_c	micro-F1 ↑	(aim: 30%)	F1 ↑	mRP ↑
0	**73.2 ± 0.5**	31.7 ± 1.1	**35.9 ± 4.7**	**39.0 ± 4.9**
0.01	**73.2 ± 0.8**	31.3 ± 2.5	30.9 ± 6.9	34.1 ± 6.3
0.1	72.8 ± 0.5	49.4 ± 20.6	26.1 ± 8.2	23.9 ± 1.6

Table 10: Development results varying λ_c (*continuity*).

In Section 5.6, we reported rationale quality on a subset of test data that includes silver and gold allegation rationales. For completeness, in Table 11 we report results on the full set of test data for silver rationales. We observe that all findings and particularly the ranking of the methods with respect to the subset of silver and gold rationales hold. Furthermore, we observe that the rationale quality performance on the full test set is slightly inferior in most cases (2-4%), which is expected as the sample annotated by the expert included only cases with allegations for articles that are more explainable.

Method	Silver rationales (31%)	
	mRP ↑	F1 ↑
RANDOM	30.7 ± 0.7	26.2 ± 0.5
HIERBERT-HA + L_s (Eq. 1)	39.0 ± 3.9	35.1 ± 3.7
HIERBERT-HA + L_s + L_g (Eq. 3)	39.1 ± 5.6	34.7 ± 5.7
HIERBERT-HA + L_s + L_g (Eq. 5)	42.7 ± 1.8	38.2 ± 1.5
HIERBERT-HA + L_s + L_r (Eq. 6)	**43.3 ± 2.3**	**39.0 ± 2.1**

Table 11: Results on rationale quality on the full set of test data for silver rationales.

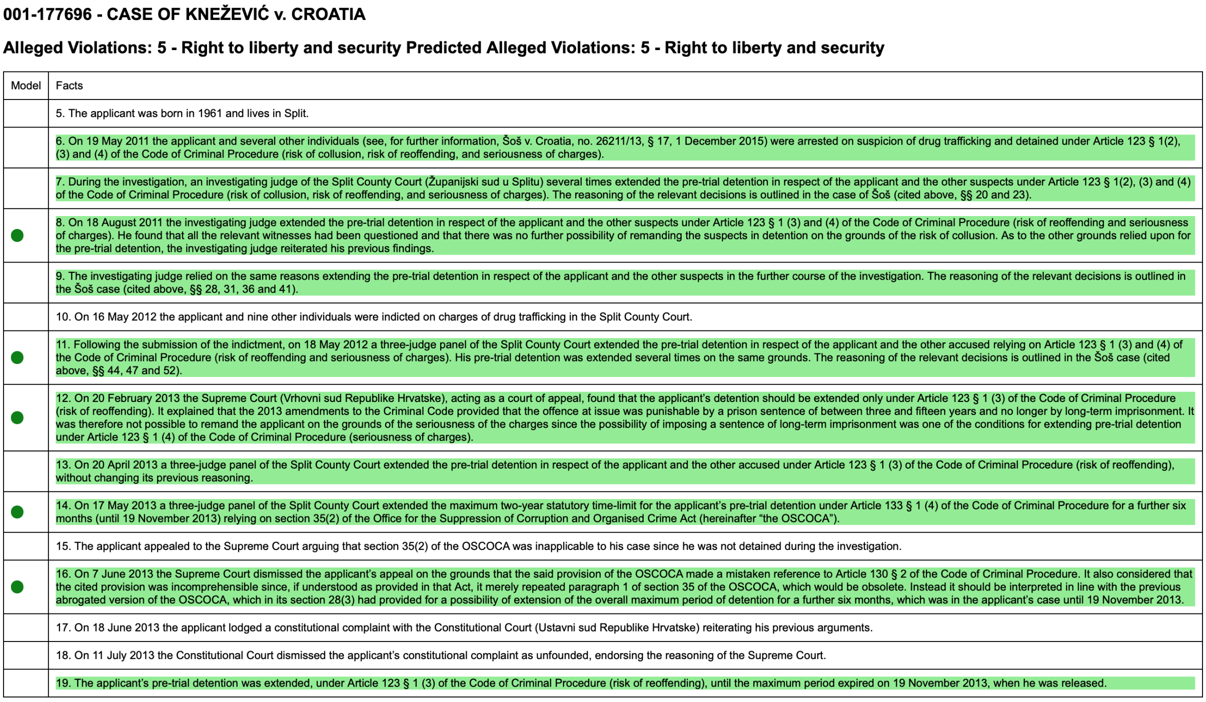
E Using ECtHR dataset via 🤗

The dataset is available at https://archive.org/details/ECtHR-NAACL2021; but you can easily load and use it in python with two lines of code:

```
from datasets import load_dataset
dataset = load_dataset("ecthr_cases")
```

F Examples of extracted Rationales from ECtHR cases with comments

In Fig. 3–7, we present examples of ECtHR cases. The highlighting (green background colour) indicates gold rationales. The dots (green dot on the left) indicate rationales extracted by our best model, HIERBERT-HA + L_s + L_g (Eq. 4, 6). In the caption of each figure, we include short comments explaining false positives (paragraphs the model wrongly selected) and false negatives (paragraphs the model wrongly missed).

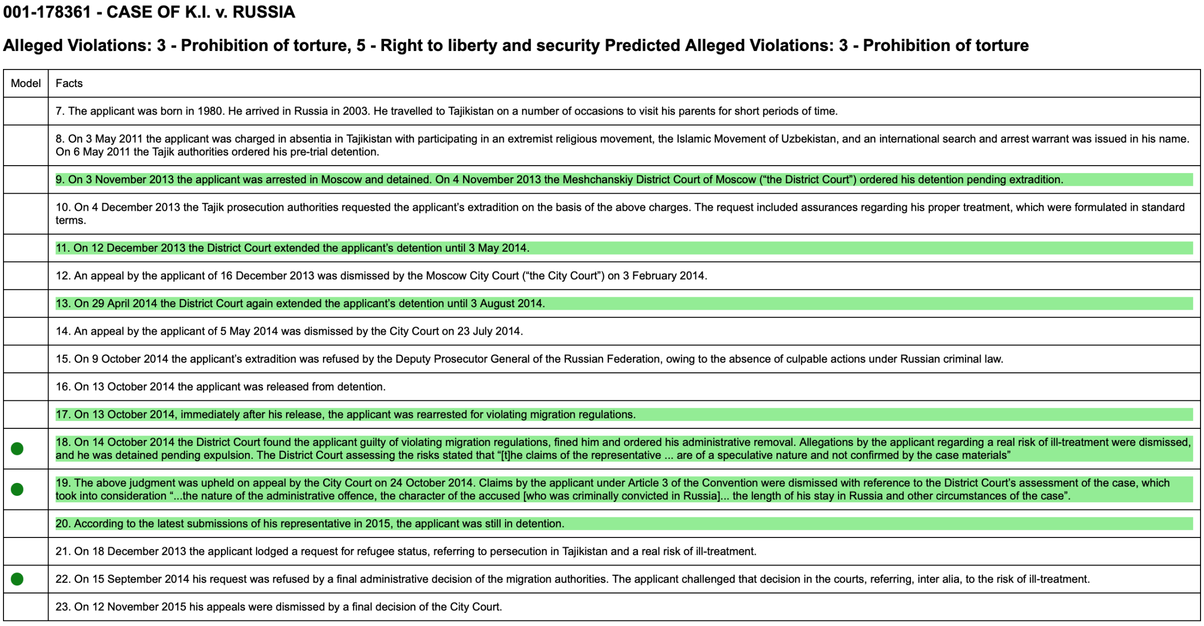


001-177696 - CASE OF KNEŽEVIĆ v. CROATIA

Alleged Violations: 5 - Right to liberty and security Predicted Alleged Violations: 5 - Right to liberty and security

Model	Facts
	5. The applicant was born in 1961 and lives in Split.
	6. On 19 May 2011 the applicant and several other individuals (see, for further information, Šoš v. Croatia, no. 26211/13, § 17, 1 December 2015) were arrested on suspicion of drug trafficking and detained under Article 123 § 1(2), (3) and (4) of the Code of Criminal Procedure (risk of collusion, risk of reoffending, and seriousness of charges).
	7. During the investigation, an investigating judge of the Split County Court (Županijski sud u Splitu) several times extended the pre-trial detention in respect of the applicant and the other suspects under Article 123 § 1(2), (3) and (4) of the Code of Criminal Procedure (risk of collusion, risk of reoffending, and seriousness of charges). The reasoning of the relevant decisions is outlined in the case of Šoš (cited above, §§ 20 and 23).
●	8. On 18 August 2011 the investigating judge extended the pre-trial detention in respect of the applicant and the other suspects under Article 123 § 1 (3) and (4) of the Code of Criminal Procedure (risk of reoffending and seriousness of charges). He found that all the relevant witnesses had been questioned and that there was no further possibility of remanding the suspects in detention on the grounds of the risk of collusion. As to the other grounds relied upon for the pre-trial detention, the investigating judge reiterated his previous findings.
	9. The investigating judge relied on the same reasons extending the pre-trial detention in respect of the applicant and the other suspects in the further course of the investigation. The reasoning of the relevant decisions is outlined in the Šoš case (cited above, §§ 28, 31, 36 and 41).
	10. On 16 May 2012 the applicant and nine other individuals were indicted on charges of drug trafficking in the Split County Court.
●	11. Following the submission of the indictment, on 18 May 2012 a three-judge panel of the Split County Court extended the pre-trial detention in respect of the applicant and the other accused relying on Article 123 § 1 (3) and (4) of the Code of Criminal Procedure (risk of reoffending and seriousness of charges). His pre-trial detention was extended several times on the same grounds. The reasoning of the relevant decisions is outlined in the Šoš case (cited above, §§ 44, 47 and 52).
●	12. On 20 February 2013 the Supreme Court (Vrhovni sud Republike Hrvatske), acting as a court of appeal, found that the applicant's detention should be extended only under Article 123 § 1 (3) of the Code of Criminal Procedure (risk of reoffending). It explained that the 2013 amendments to the Criminal Code provided that the offence at issue was punishable by a prison sentence of between three and fifteen years and no longer by long-term imprisonment. It was therefore not possible to remand the applicant on the grounds of the seriousness of the charges since the possibility of imposing a sentence of long-term imprisonment was one of the conditions for extending pre-trial detention under Article 123 § 1 (4) of the Code of Criminal Procedure (seriousness of charges).
	13. On 20 April 2013 a three-judge panel of the Split County Court extended the pre-trial detention in respect of the applicant and the other accused under Article 123 § 1 (3) of the Code of Criminal Procedure (risk of reoffending), without changing its previous reasoning.
●	14. On 17 May 2013 a three-judge panel of the Split County Court extended the maximum two-year statutory time-limit for the applicant's pre-trial detention under Article 133 § 1 (4) of the Code of Criminal Procedure for a further six months (until 19 November 2013) relying on section 35(2) of the Office for the Suppression of Corruption and Organised Crime Act (hereinafter "the OSCOCA").
	15. The applicant appealed to the Supreme Court arguing that section 35(2) of the OSCOCA was inapplicable to his case since he was not detained during the investigation.
●	16. On 7 June 2013 the Supreme Court dismissed the applicant's appeal on the grounds that the said provision of the OSCOCA made a mistaken reference to Article 130 § 2 of the Code of Criminal Procedure. It also considered that the cited provision was incomprehensible since, if understood as provided in that Act, it merely repeated paragraph 1 of section 35 of the OSCOCA, which would be obsolete. Instead it should be interpreted in line with the previous abrogated version of the OSCOCA, which in its section 28(3) had provided for a possibility of extension of the overall maximum period of detention for a further six months, which was in the applicant's case until 19 November 2013.
	17. On 18 June 2013 the applicant lodged a constitutional complaint with the Constitutional Court (Ustavni sud Republike Hrvatske) reiterating his previous arguments.
	18. On 11 July 2013 the Constitutional Court dismissed the applicant's constitutional complaint as unfounded, endorsing the reasoning of the Supreme Court.
	19. The applicant's pre-trial detention was extended, under Article 123 § 1 (3) of the Code of Criminal Procedure (risk of reoffending), until the maximum period expired on 19 November 2013, when he was released.

Figure 3: (KNEZEVIC v. CROATIA, No. 55133/13}) The model extracted most of the relevant facts indicating a possible violation of Article 5. Note that 67% (10 of 15) of the facts were considered relevant by the legal expert. Our model has a disadvantage in this case because, being trained to operate at a predefined sparsity level (30%), it extracted only 5 of the 15 facts (33%).

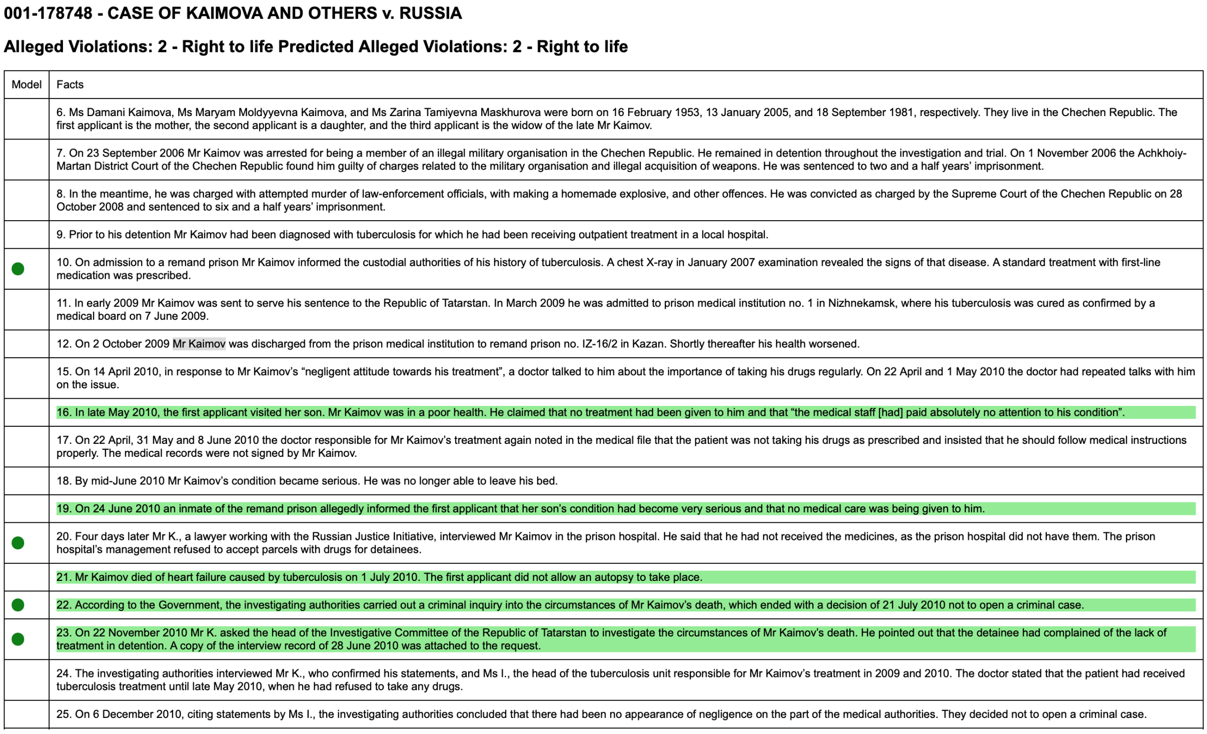


001-178361 - CASE OF K.I. v. RUSSIA

Alleged Violations: 3 - Prohibition of torture, 5 - Right to liberty and security Predicted Alleged Violations: 3 - Prohibition of torture

Model	Facts
	7. The applicant was born in 1980. He arrived in Russia in 2003. He travelled to Tajikistan on a number of occasions to visit his parents for short periods of time.
	8. On 3 May 2011 the applicant was charged in absentia in Tajikistan with participating in an extremist religious movement, the Islamic Movement of Uzbekistan, and an international search and arrest warrant was issued in his name. On 6 May 2011 the Tajik authorities ordered his pre-trial detention.
	9. On 3 November 2013 the applicant was arrested in Moscow and detained. On 4 November 2013 the Meshchanskiy District Court of Moscow ("the District Court") ordered his detention pending extradition.
	10. On 4 December 2013 the Tajik prosecution authorities requested the applicant's extradition on the basis of the above charges. The request included assurances regarding his proper treatment, which were formulated in standard terms.
	11. On 12 December 2013 the District Court extended the applicant's detention until 3 May 2014.
	12. An appeal by the applicant of 16 December 2013 was dismissed by the Moscow City Court ("the City Court") on 3 February 2014.
	13. On 29 April 2014 the District Court again extended the applicant's detention until 3 August 2014.
	14. An appeal by the applicant of 5 May 2014 was dismissed by the City Court on 23 July 2014.
	15. On 9 October 2014 the applicant's extradition was refused by the Deputy Prosecutor General of the Russian Federation, owing to the absence of culpable actions under Russian criminal law.
	16. On 13 October 2014 the applicant was released from detention.
	17. On 13 October 2014, immediately after his release, the applicant was rearrested for violating migration regulations.
●	18. On 14 October 2014 the District Court found the applicant guilty of violating migration regulations, fined him and ordered his administrative removal. Allegations by the applicant regarding a real risk of ill-treatment were dismissed, and he was detained pending expulsion. The District Court assessing the risks stated that "[t]he claims of the representative ... are of a speculative nature and not confirmed by the case materials"
●	19. The above judgment was upheld on appeal by the City Court on 24 October 2014. Claims by the applicant under Article 3 of the Convention were dismissed with reference to the District Court's assessment of the case, which took into consideration "...the nature of the administrative offence, the character of the accused [who was criminally convicted in Russia]... the length of his stay in Russia and other circumstances of the case".
	20. According to the latest submissions of his representative in 2015, the applicant was still in detention.
	21. On 18 December 2013 the applicant lodged a request for refugee status, referring to persecution in Tajikistan and a real risk of ill-treatment.
●	22. On 15 September 2014 his request was refused by a final administrative decision of the migration authorities. The applicant challenged that decision in the courts, referring, inter alia, to the risk of ill-treatment.
	23. On 12 November 2015 his appeals were dismissed by a final decision of the City Court.

Figure 4: (K.I. v. RUSSIA, No. 58182/14) Paragraphs 9, 11, 13 and 20 clearly indicate plausible violation of the right to liberty (Article 5), as they refer to continuous extension of applicant detention, but our model was unable to extract them, thus it was unable to predict this allegation. The model targeted only paragraphs that indicate ill-treatment, which is connected to plausible violation of Article 3 (Prohibition of Torture).

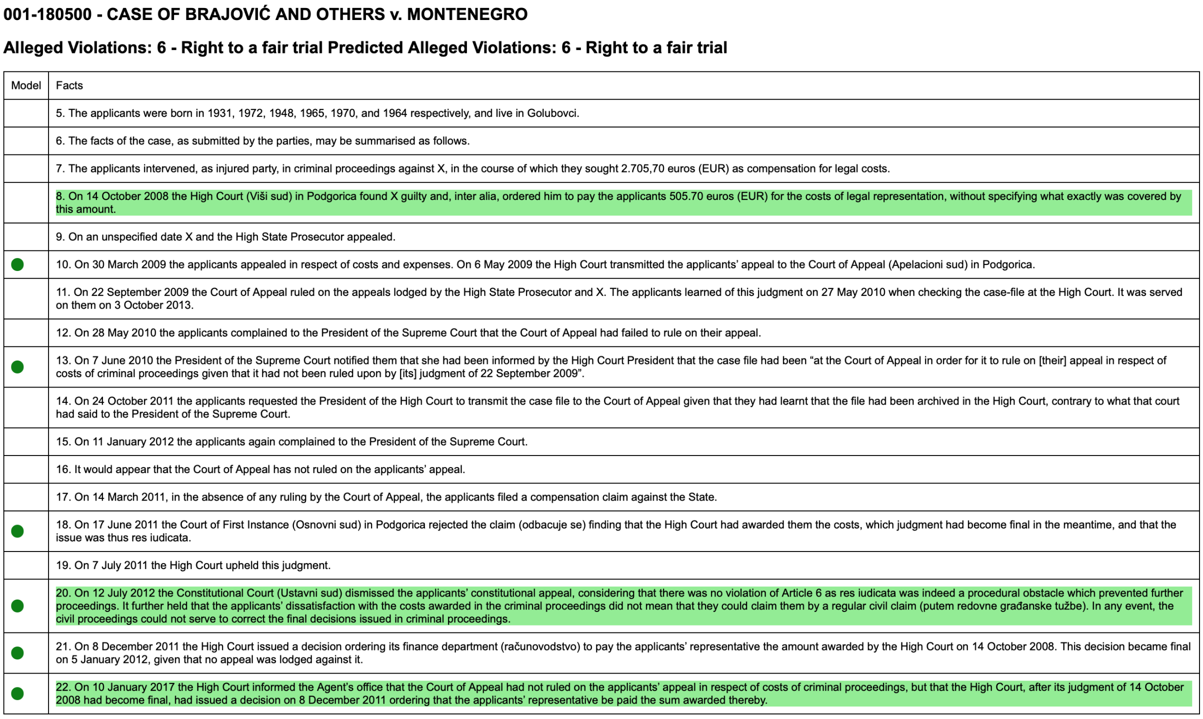


001-178748 - CASE OF KAIMOVA AND OTHERS v. RUSSIA

Alleged Violations: 2 - Right to life Predicted Alleged Violations: 2 - Right to life

Model	Facts
	6. Ms Damani Kaimova, Ms Maryam Moldyyevna Kaimova, and Ms Zarina Tamiyevna Maskhurova were born on 16 February 1953, 13 January 2005, and 18 September 1981, respectively. They live in the Chechen Republic. The first applicant is the mother, the second applicant is a daughter, and the third applicant is the widow of the late Mr Kaimov.
	7. On 23 September 2006 Mr Kaimov was arrested for being a member of an illegal military organisation in the Chechen Republic. He remained in detention throughout the investigation and trial. On 1 November 2006 the Achkhoiy-Martan District Court of the Chechen Republic found him guilty of charges related to the military organisation and illegal acquisition of weapons. He was sentenced to two and a half years' imprisonment.
	8. In the meantime, he was charged with attempted murder of law-enforcement officials, with making a homemade explosive, and other offences. He was convicted as charged by the Supreme Court of the Chechen Republic on 28 October 2008 and sentenced to six and a half years' imprisonment.
	9. Prior to his detention Mr Kaimov had been diagnosed with tuberculosis for which he had been receiving outpatient treatment in a local hospital.
●	10. On admission to a remand prison Mr Kaimov informed the custodial authorities of his history of tuberculosis. A chest X-ray in January 2007 examination revealed the signs of that disease. A standard treatment with first-line medication was prescribed.
	11. In early 2009 Mr Kaimov was sent to serve his sentence to the Republic of Tatarstan. In March 2009 he was admitted to prison medical institution no. 1 in Nizhnekamsk, where his tuberculosis was cured as confirmed by a medical board on 7 June 2009.
	12. On 2 October 2009 Mr Kaimov was discharged from the prison medical institution to remand prison no. IZ-16/2 in Kazan. Shortly thereafter his health worsened.
	15. On 14 April 2010, in response to Mr Kaimov's "negligent attitude towards his treatment", a doctor talked to him about the importance of taking his drugs regularly. On 22 April and 1 May 2010 the doctor had repeated talks with him on the issue.
	16. In late May 2010, the first applicant visited her son. Mr Kaimov was in a poor health. He claimed that no treatment had been given to him and that "the medical staff [had] paid absolutely no attention to his condition".
	17. On 22 April, 31 May and 8 June 2010 the doctor responsible for Mr Kaimov's treatment again noted in the medical file that the patient was not taking his drugs as prescribed and insisted that he should follow medical instructions properly. The medical records were not signed by Mr Kaimov.
	18. By mid-June 2010 Mr Kaimov's condition became serious. He was no longer able to leave his bed.
	19. On 24 June 2010 an inmate of the remand prison allegedly informed the first applicant that her son's condition had become very serious and that no medical care was being given to him.
●	20. Four days later Mr K., a lawyer working with the Russian Justice Initiative, interviewed Mr Kaimov in the prison hospital. He said that he had not received the medicines, as the prison hospital did not have them. The prison hospital's management refused to accept parcels with drugs for detainees.
	21. Mr Kaimov died of heart failure caused by tuberculosis on 1 July 2010. The first applicant did not allow an autopsy to take place.
●	22. According to the Government, the investigating authorities carried out a criminal inquiry into the circumstances of Mr Kaimov's death, which ended with a decision of 21 July 2010 not to open a criminal case.
●	23. On 22 November 2010 Mr K. asked the head of the Investigative Committee of the Republic of Tatarstan to investigate the circumstances of Mr Kaimov's death. He pointed out that the detainee had complained of the lack of treatment in detention. A copy of the interview record of 28 June 2010 was attached to the request.
	24. The investigating authorities interviewed Mr K., who confirmed his statements, and Ms I., the head of the tuberculosis unit responsible for Mr Kaimov's treatment in 2009 and 2010. The doctor stated that the patient had received tuberculosis treatment until late May 2010, when he had refused to take any drugs.
	25. On 6 December 2010, citing statements by Ms I., the investigating authorities concluded that there had been no appearance of negligence on the part of the medical authorities. They decided not to open a criminal case.

Figure 5: (KAIMOVA AND OTHERS v. RUSSIA, No. 58182/14) Paragraphs 16 and 19 clearly indicate that the applicant's health (life) was at risk and authorities did not pay attention, but these paragraphs were not selected by the model. Instead paragraph 10 states that the applicant initially informed the authorities for his medical history and they provided medication. This is an indication of model sensitivity to language describing health issues (tuberculosis) in general and not specific well-defined allegations for ill-treatment on the merits.

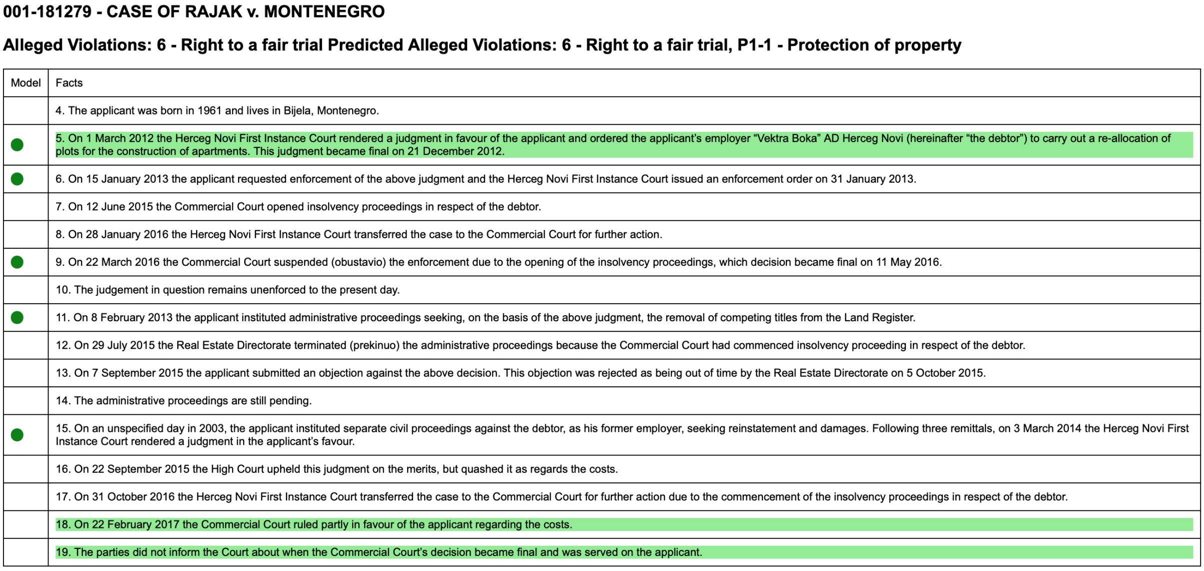


001-180500 - CASE OF BRAJOVIĆ AND OTHERS v. MONTENEGRO

Alleged Violations: 6 - Right to a fair trial Predicted Alleged Violations: 6 - Right to a fair trial

Model	Facts
	5. The applicants were born in 1931, 1972, 1948, 1965, 1970, and 1964 respectively, and live in Golubovci.
	6. The facts of the case, as submitted by the parties, may be summarised as follows.
	7. The applicants intervened, as injured party, in criminal proceedings against X, in the course of which they sought 2.705,70 euros (EUR) as compensation for legal costs.
	8. On 14 October 2008 the High Court (Viši sud) in Podgorica found X guilty and, inter alia, ordered him to pay the applicants 505.70 euros (EUR) for the costs of legal representation, without specifying what exactly was covered by this amount.
	9. On an unspecified date X and the High State Prosecutor appealed.
●	10. On 30 March 2009 the applicants appealed in respect of costs and expenses. On 6 May 2009 the High Court transmitted the applicants' appeal to the Court of Appeal (Apelacioni sud) in Podgorica.
	11. On 22 September 2009 the Court of Appeal ruled on the appeals lodged by the High State Prosecutor and X. The applicants learned of this judgment on 27 May 2010 when checking the case-file at the High Court. It was served on them on 3 October 2013.
	12. On 28 May 2010 the applicants complained to the President of the Supreme Court that the Court of Appeal had failed to rule on their appeal.
●	13. On 7 June 2010 the President of the Supreme Court notified them that she had been informed by the High Court President that the case file had been "at the Court of Appeal in order for it to rule on [their] appeal in respect of costs of criminal proceedings given that it had not been ruled upon by [its] judgment of 22 September 2009".
	14. On 24 October 2011 the applicants requested the President of the High Court to transmit the case file to the Court of Appeal given that they had learnt that the file had been archived in the High Court, contrary to what that court had said to the President of the Supreme Court.
	15. On 11 January 2012 the applicants again complained to the President of the Supreme Court.
	16. It would appear that the Court of Appeal has not ruled on the applicants' appeal.
	17. On 14 March 2011, in the absence of any ruling by the Court of Appeal, the applicants filed a compensation claim against the State.
●	18. On 17 June 2011 the Court of First Instance (Osnovni sud) in Podgorica rejected the claim (odbacuje se) finding that the High Court had awarded them the costs, which judgment had become final in the meantime, and that the issue was thus res iudicata.
	19. On 7 July 2011 the High Court upheld this judgment.
●	20. On 12 July 2012 the Constitutional Court (Ustavni sud) dismissed the applicants' constitutional appeal, considering that there was no violation of Article 6 as res iudicata was indeed a procedural obstacle which prevented further proceedings. It further held that the applicants' dissatisfaction with the costs awarded in the criminal proceedings did not mean that they could claim them by a regular civil claim (putem redovne građanske tužbe). In any event, the civil proceedings could not serve to correct the final decisions issued in criminal proceedings.
●	21. On 8 December 2011 the High Court issued a decision ordering its finance department (računovodstvo) to pay the applicants' representative the amount awarded by the High Court on 14 October 2008. This decision became final on 5 January 2012, given that no appeal was lodged against it.
●	22. On 10 January 2017 the High Court informed the Agent's office that the Court of Appeal had not ruled on the applicants' appeal in respect of costs of criminal proceedings, but that the High Court, after its judgment of 14 October 2008 had become final, had issued a decision on 8 December 2011 ordering that the applicants' representative be paid the sum awarded thereby.

Figure 6: (BRAJOVIC AND OTHERS v. MONTENEGRO, No. 52529/12) A causal inference would connect paragraph 8 (initial trial) with paragraphs 20–22 (next trials) to infer mistrial, because there was is verdict after a reasonable period of time. Instead the model seems to be sensitive to references for the involvement of higher courts as justification of mistrial (paragraphs 10, 13, 18, and 21). This suggests that the model probably follows poor (greedy) reasoning, i.e., if the applicant appealed to higher courts, then the case is mistreated.



001-181279 - CASE OF RAJAK v. MONTENEGRO

Alleged Violations: 6 - Right to a fair trial Predicted Alleged Violations: 6 - Right to a fair trial, P1-1 - Protection of property

Model	Facts
	4. The applicant was born in 1961 and lives in Bijela, Montenegro.
●	5. On 1 March 2012 the Herceg Novi First Instance Court rendered a judgment in favour of the applicant and ordered the applicant's employer "Vektra Boka" AD Herceg Novi (hereinafter "the debtor") to carry out a re-allocation of plots for the construction of apartments. This judgment became final on 21 December 2012.
●	6. On 15 January 2013 the applicant requested enforcement of the above judgment and the Herceg Novi First Instance Court issued an enforcement order on 31 January 2013.
	7. On 12 June 2015 the Commercial Court opened insolvency proceedings in respect of the debtor.
	8. On 28 January 2016 the Herceg Novi First Instance Court transferred the case to the Commercial Court for further action.
●	9. On 22 March 2016 the Commercial Court suspended (obustavio) the enforcement due to the opening of the insolvency proceedings, which decision became final on 11 May 2016.
	10. The judgement in question remains unenforced to the present day.
●	11. On 8 February 2013 the applicant instituted administrative proceedings seeking, on the basis of the above judgment, the removal of competing titles from the Land Register.
	12. On 29 July 2015 the Real Estate Directorate terminated (prekinuo) the administrative proceedings because the Commercial Court had commenced insolvency proceeding in respect of the debtor.
	13. On 7 September 2015 the applicant submitted an objection against the above decision. This objection was rejected as being out of time by the Real Estate Directorate on 5 October 2015.
	14. The administrative proceedings are still pending.
●	15. On an unspecified day in 2003, the applicant instituted separate civil proceedings against the debtor, as his former employer, seeking reinstatement and damages. Following three remittals, on 3 March 2014 the Herceg Novi First Instance Court rendered a judgment in the applicant's favour.
	16. On 22 September 2015 the High Court upheld this judgment on the merits, but quashed it as regards the costs.
	17. On 31 October 2016 the Herceg Novi First Instance Court transferred the case to the Commercial Court for further action due to the commencement of the insolvency proceedings in respect of the debtor.
	18. On 22 February 2017 the Commercial Court ruled partly in favour of the applicant regarding the costs.
	19. The parties did not inform the Court about when the Commercial Court's decision became final and was served on the applicant.

Figure 7: (RAJAK v. MONTENEGRO, No. 71998/11) Similarly to the case presented in Fig. 6, the main argument in this case is mistrial because there was a verdict after a reasonable period of time (paragraphs 5 and 18-19). The model selected paragraph 11, which does not indicate plausible violations. Given the model's prediction for allegations with respect to Article 1 of the 1st Protocol on the protection of property, we believe that paragraph 11 was selected by the model as justification on that matter.