From RoBERTa to aLEXa:
Automated Legal Expert Arbitrator
for Neural Legal Judgment Prediction

Stanford CS224N Custom Project Final Report

Lukas Haas
Department of Computer Science
Stanford University
lukashaas@stanford.edu

Michal Skreta
Department of Computer Science
Stanford University
michal.skreta@stanford.edu

Abstract

Neural models have shown promise in redefining the field of legal judgment prediction (LJP), serving as an aid for judges while helping citizens assess the fairness of judgments. Previous neural LJP approaches on binary and multi-label classification tasks, however, relied on outdated language models and faulty hyperparameters. Beyond improving state-of-the-art results in both a binary and the multi-label setting by fine-tuning large language models, we introduce aLEXa, a multi-task hierarchical language model with self-learning loss weights and attention forcing, teaching the model what legal facts to pay most attention to. Additionally, we improve the explainability of LJP models through paragraph attention weighting visualizations, allowing us to qualitatively assess the quality of legal predictions in addition to traditional quantitative metrics. Our results underscore the potential of NLP approaches to redefine traditional judicial decision-making and show promise of the efficacy of hierarchical and domain fine-tuned language models.

1   Key Information

- **Mentor**: Lucia Zheng (zlucia@stanford.edu).
- **External Collaborators**: None.
- **Sharing project**: No.

2   Introduction

Motivation. For as long as laws existed, they have been largely written in a textual form. Consequently, natural language processing (NLP) techniques have naturally shown promise as a revolutionizing force in the field of legal analysis, as noted by Nallapati and Manning in 2008 [1]. One particular application of NLP in law, the one studied in this paper, is legal judgment prediction (LJP). The goal of LJP is to predict a case's outcome based on text describing facts of a legal case. This task is of enormous societal importance: not only does it provide a useful reference to the judges, but it also helps regular citizens by reducing legal costs and aids human rights organizations in better assessing the fairness of the judgments.

Shortcomings of existing approaches. Although legal cases are usually represented in textual form, computational analysis has not been widely implemented in legal judgment prediction. Before the Chalkidis et al. (2019) paper [2] was published (described in Section 3), previous publications that considered LJP in English have focused on linear models with features based on bags of words and topics to represent legal textual information extracted from cases [3]. More sophisticated neural models have been used in the field, but only in Chinese [4]. The datasets have also suffered from a

small size and limited richness, hampering the efficacy of neural methods. Even as research in NLP started to make waves, most papers have treated domain-agnostic rather than domain-specific topics, thus providing insufficient attention to fields such as neural LJP.

Our contribution. We go beyond the prior approaches to neural legal judgment prediction by building transformer-based neural networks and achieving state-of-the-art results on binary and multi-label classification problems in the field of legal judgment prediction, uncovering the potential of NLP to serve as an aid for judges while helping citizens assess the fairness of judgments. As part of our work, we propose novel hierarchical network architectures in a multi-task setting showing great promise in both performance and explainability to generate decision rationales based on case facts.

3 Related Work

The key academic paper we use as a benchmark throughout our work is "Neural Legal Judgment Prediction in English" by Chalkidis et al. (2019) [2].

The authors of the paper go beyond traditional, hand-crafted methods employed in legal NLP in English thus far, and use end-to-end neural models to improve the state-of-the-art models in predicting the outcomes of legal cases. The main contribution of the paper is the presentation of a novel dataset of approximately 11.5k cases with outcomes from the European Court of Human Rights (ECHR), vastly expanding the availability of large legal corpora compared to the previous standards in the field [3]. Additionally, and perhaps more importantly, the authors use their novel dataset to evaluate a range of neural models in English, which is done for the first time in the discipline of legal NLP. They run the models on three distinct tasks; not only do they focus on traditional binary classification (whether or not a violation of an Article of the European Convention on Human Rights occurred), but they also zero in on multi-label classification (determining the precise type of violation) as well as on detecting the importance of any given case as labelled by experts. Their hierarchical BERT model becomes the state of the art for both binary classification with an F1 score of 82.0 and multi-label classification with an F1 score of 60.0 [2]. Despite these advances, their BERT model performs worse than random and their experimental specifications hinder the frontier of the state of the art in neural judgment prediction, a fact that we exploit in the following sections.

4 Approach

Tasks. Our task for model development and evaluation is two-fold, consisting of (1) a binary classification and (2) a multi-class classification task. Given a list of paragraphs containing the facts of the case, our goal is to (1) predict whether a human rights violation was found by the court, and (2) which human rights article was violated, if any. Task (2) is especially challenging as the specific human rights articles that are violated do not occur with equal frequency in the dataset; 11 out of 66 article labels occur less than 50 times as later presented in Figure 3 in Section 5.1.

Methods. In approaching our binary and multi-label tasks of human rights article violation classification, we first build a variety of baselines using pre-trained large language models (LLMs). In order to address the limitation of these models only processing a limited number of tokens, we then proceed to building hierarchical language models which first embed paragraph knowledge and then produce an aggregate case embedding used for classification. Finally, we introduce **A**utomated **L**egal **Ex**pert **A**rbitrator (aLEXa), a multi-task hierarchical language model which simultaneously learns to predict the case outcomes and to select the relevant case background facts as a judgment rationale.

Baselines. For our baselines, we use already pre-trained versions of **BERT** [5], **RoBERTa** [6], and **LEGAL-BERT** [7] models available on Hugging Face to then fine-tune on our classification (binary and multi-label) downstream tasks. Our rationale behind this decision was that these models each provide a great baseline, for different reasons: BERT is the original and most used transformer-based NLP model, RoBERTa further outperforms BERT by training for more epochs on more data while slightly changing the pre-training objective, and finally LEGAL-BERT, a BERT model completely trained on legal text corpora. Although we experimented with the more expressive RoBERTa large model from Hugging Face [6], we decided against including that model in our approach as we ran into GPU memory problems even with a batch size of only a single sample. The limited computational resources available to us on Microsoft Azure further encouraged us to focus on the three former

pre-trained models. Since BERT, RoBERTa, and LEGAL-BERT all have a word token limit of 512 tokens, we trained our baselines only on the first 512 tokens of every case.

Hierarchical Large Language Models (HIER-LLMs). Traditionally, applications of NLP in the legal domain have struggled with modelling complex dependencies often stretching over many paragraphs. In Chalkidis et al. (2019) [2], the authors address this limitation explicitly by explaining how commonly used LLMs struggle at "*processing long documents*" due to their fixed word token length.

To address this limitation and go beyond the 512 word token limit, we build a hierarchical LLM architecture which captures long-term dependencies and can use any BERT-based model as a base model. Our work was inspired by the Lu et al. 2021 paper [8] which introduces a sentence-level hierarchical BERT model for document classification. As can be seen in Figure 1, our proposed model architecture first feeds every paragraph of the legal case through a base model (i.e RoBERTa) to obtain paragraph-level embeddings. In order to model long dependencies across paragraphs, the paragraphs embeddings are again fed through two transformer layers which contain multi-head self-attention mechanisms to compare paragraphs with each other. The transformer layers then output an aggregate case embedding which is a compact vector representation of all relevant facts in the case which is then fed through a dropout and linear classification layer to obtain predictions for either the binary or multi-label tasks.
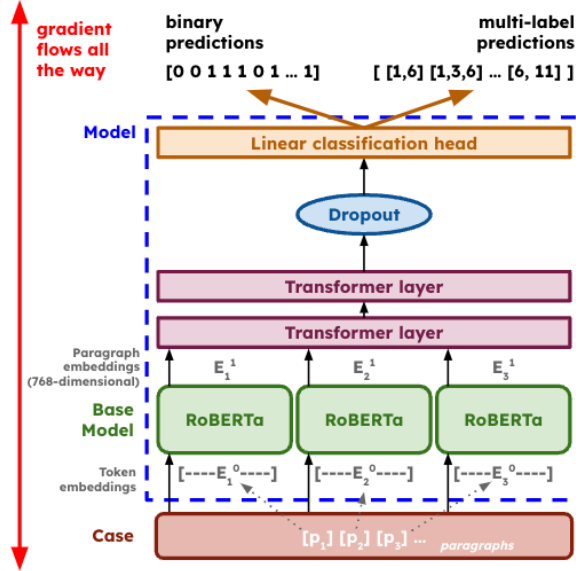

Figure 1: Hierarchical LLM model architecture (base model can vary).

A major challenge in our work was to find an efficient method to combine all legal paragraph embeddings into a final case embedding while also generating attention scores over paragraphs which would make the model more interpretable. As part of our work, we experiment with both a simple multi-head self-attention layer as well as transformer encoder layers and find that given our dataset, two transformer layers generate the best results. Every transformer layer consists of a multi-head self-attention layer followed by two fully connected layers, all of which have skip connections, dropout, and batch normalization layers between them.

A limitation of Lu et al. (2021) [8] is that the authors assume that LLMs such as RoBERTa can already encode token-level information well and thus the base model does not need to be fine-tuned during training. To address this limitation and leverage the benefits of fine-tuning LLMs on downstream tasks, our hierarchical model can be trained in an end-to-end fashion with gradients being back-propagated all the way from the case embeddings to the token-level embeddings.

Automated Legal Expert Arbitrator (aLEXa). A second limitation voiced in Chalkidis et al. (2019) [2] is that neural judgement prediction models struggle to give a justification for their predictions.

Because of this, we expand on our Hierarchical LLM model architecture to generate salient judgement rationales as an explicit part of the training procedure.

We introduce **A**utomated **L**egal **Ex**pert **A**rbitrator (aLEXa), a multi-task hierarchical language model which explicitly is taught what paragraphs to pay attention to via attention forcing inspired by Dou et al. (2021) [9]. This allows aLEXa to simultaneously predict the case outcome (binary or multi-label classification) while also predicting which paragraphs of background case facts are relevant for the final court decision. As can be seen in Figure 2, similar to the HIER-LLM model architecture, aLEXa's processes each case paragraph separately using a base LLM to produce separate paragraph embeddings. These paragraph embeddings are then fed through a single transformer encoder layer to produce an aggregate case embedding which again is used for classification. We use a single transformer layer due to attention scores being more attributable to specific paragraphs as opposed to when using multiple, stacked transformer layers.
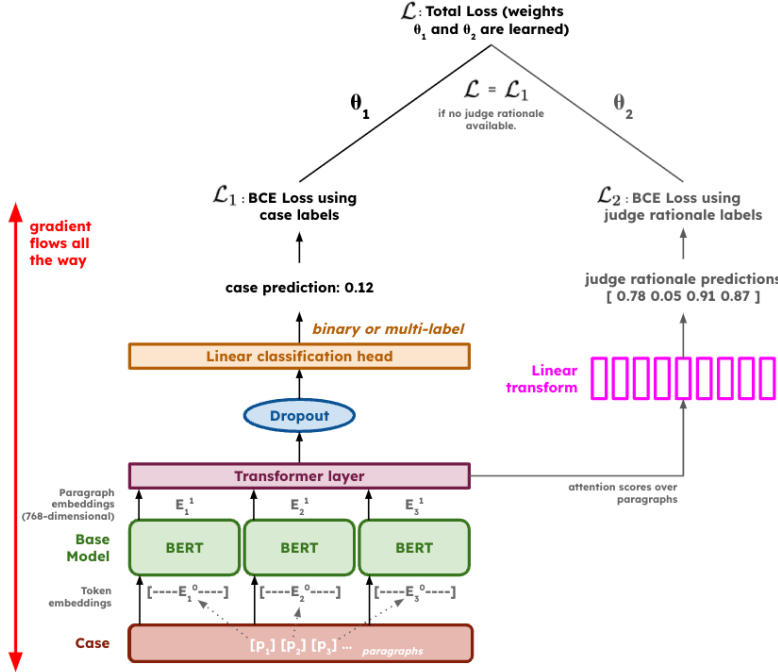

Figure 2: aLEXa model architecture (base model can vary).

Our aLEXa model architecture introduces two novel contributions to the field of neural legal judgement prediction. First, as can be seen in Figure 2, paragraph attention scores from the multi-head self-attention mechanism in the transformer layer are extracted and linearly transformed (same transformation for each paragraph) to produce binary relevance predictions for each paragraph. This not only allows the model to directly output a decision rationale but also provides a mechanism to explicitly teach the model which paragraphs to pay attention during training to via a multi-label paragraph classification task. For teaching paragraph relevance, we use the augmented dataset by Chalkidis et al. (2021) [10].

The second contribution of aLEXa is that the model is trained in a multi-task setting with self-learning loss weights between the neural judgement classification task (weight σ_1 for loss $\mathcal{L}_1$) and the attention forcing multi-label classification task (weight σ_2 for loss $\mathcal{L}_2$). The reason why we chose to let aLEXa learn the weights between both losses itself was that for 50.2% of training examples, judgment rationales (paragraph relevance labels) were not available, varying between batches, and thus learning the loss weights directly proves to be more stable during training. To also account for examples which did not have judgement rationales, we create an indicator variable $a^{(i)}$, equal to one if the paragraph labels were present, zero otherwise. Our final loss function is an adaptation of the

multi-task homoscedastic uncertainty loss function mentioned in Kendall et al. (2017) [11] which is shown in Equation 1 for a single sample (i).

$$\mathcal{L}^{(i)} = a^{(i)} \left(\frac{1}{2\sigma_1^2} \mathcal{L}_1^{(i)} + \frac{1}{2\sigma_2^2} \mathcal{L}_2^{(i)} + \log \sigma_1 \sigma_2 \right) + (1 - a^{(i)})\mathcal{L}_1^{(i)} \qquad (1)$$

5 Experiments

5.1 Data

We focus our work on the dataset presented in the Chalkidis et al. (2019) paper which is a dataset from a publicly available database of the European Court of Human Rights (ECHR) consisting of 11,478 cases with associated outcomes [12]. For each case, the dataset contains facts from the case description that were extracted using regular expressions (throughout our paper, we also refer to these "facts" as "paragraphs", in a manner that is interchangeable). The paper contains a split of 7,100/1,380/2,998 cases between the training, validation, and test sets, respectively, and we use that pre-defined split in our work.

Additionally, each case is mapped to articles of the European Convention on Human Rights that were violated, if any, with a total of 66 different article labels. The labels suffer from substantial class imbalance as 11 of these labels occur less than 50 times, and only 21 of the labels occur in the training set. The distribution of the number of violations by article based on our exploratory data analysis are visualized in Figure 3. Interestingly, Article 6 of the ECHR, the right to a fair trial, is by the most commonly seen one, followed by Article 5: the right to liberty and security. In addition, articles that we initially thought might be common, such as Article 9: freedom of thought, conscience and religion and Article 14: prohibition of discrimination, are occurring very early, suggesting a specific definition of the type of cases heard by the ECHR.

For our Automated Legal Expert Arbitrator (aLEXa), we enrich this dataset with "silver judgment rationales", or paragraph decision relevance labels, cited in the ECHR judges' rationales as presented by Chalkidis et al. (2021) [10]. This allows us to obtain decision rationales for 49.8%, 48.5% and 65.0% of training, validation, and test samples, respectively.

5.2 Evaluation method

We evaluated our models using three standard evaluation metrics in a similar fashion to Chalkidis et al. (2019) [2], i.e. precision, recall, and F1 score. This decision was motivated both by compliance with research standards and by comparability with existing literature. More specifically, we used a *macro* F1

Figure 3: Number of article violations.

score to evaluate our binary classification task, which means that we weighed the the performance of both prediction classes equally. For the multi-label classification task, however, due to the high label imbalance in article violation frequency between, we employed a *micro* F1 score (μ-F1), meaning that we weighed the performance of each class by the frequency of the corresponding class label which more closely models real-world conditions, as suggested by Figure 3.
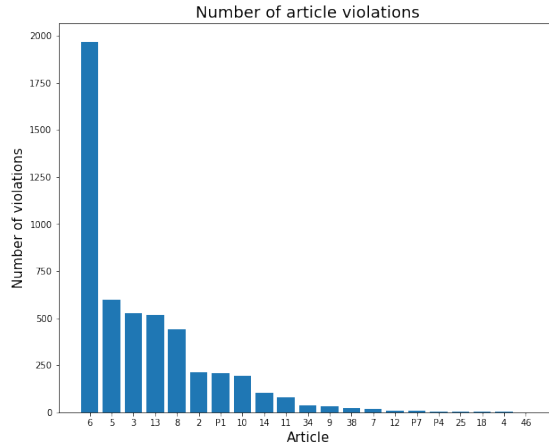
5.3 Experimental details

Due to the Microsoft Azure virtual machine GPU memory limitations, we were only able to feed in one case example at a time. Consequently, our gradient has been extremely oscillating. We managed to fix this problem by employing gradient accumulation, accumulating gradient of 64 case examples at once before updating our parameters, smoothing our training batch loss.

A key decision we needed to make regarding the parameters in our experiment concerned the number of paragraph embeddings and token embeddings used per paragraph. We multiplied these two numbers to obtain a metric proportional to the number of parameters, and estimated that a total of 10,000 parameters would be best. We subsequently created a metric for case word coverage, and plotted different combinations of parameters and coverage as seen in Figure 4. The darker values represented higher parameter values for paragraph embeddings, with each "line" being equivalent to a fixed number of tokens per paragraph. The visualization made us realize that marginal returns from larger token embeddings are more valuable than those from paragraph embeddings. Ultimately, through a grid search on data subsets we found that processing 48 paragraphs per case each with 224 tokens worked best given our limited resources while achieving a satisfactory coverage rate.

For all our models, we experimented with different learning rates through a hyperparameter grid search on a subset of all data and found that a learning rate of $\alpha = 2 \cdot 10^{-5}$ gave the best results while still allowing for fast model iteration.

As part of our experimental setting, we also experimented with the learning rate for BERT recommended in Chalkidis et al. (2019) [2] which is $\alpha = 1 \cdot 10^{-3}$. We noticed, however, that a learning rate this high caused our models' training losses to diverge, resulting in a performance inferior to a random-guess baseline. This was not addressed by Chalkidis et al. in their 2019 paper [2].
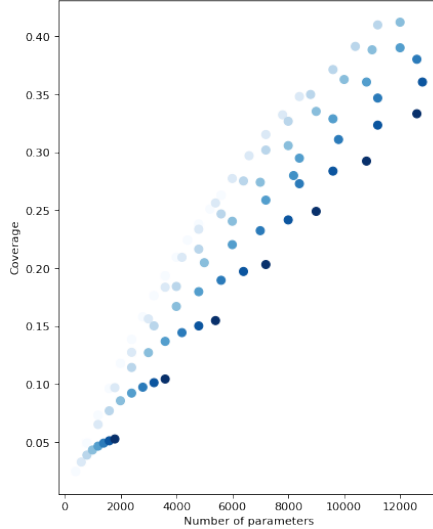

Figure 4: Finding the optimal parameter values.

5.4 Results

Discussion of quantitative results from binary classification. The binary human rights violation classification results obtained by us can be seen in Table 1. While contrasting our results with Chalkidis et al. (2019), one can see that our BERT and RoBERTa baselines already outperformed all models in Chalkidis et al. (2019) when looking at the aggregate metric of macro F1. This result was initially unexpected but after an analysis of the learning rate chosen by Chalkidis et al., we found that their learning rate of $1e^{-3}$ diverged for our models, validating our findings.

By building hierarchical LLMs aggregating paragraph embeddings into a case embedding we aimed to increase our precision by enabling the processing of longer texts which we achieved as expected. However, we did not expect the recall to drop significantly which was likely due the higher data dimensionality and model expressiveness which could in the future be regularized to potentially achieve better results. Still, the hierarchical architecture structure improved our state-of-the-art results.

Finally, by actively teaching attention scores using the aLEXa model we were able to further improve our state-of-the-art results, achieving a high score of **83.4%** with the aLEXa model using the BERT

base. While we expected attention forcing to aid in model interpretability we were delighted to see that it also improved results for case classification tasks, validating our decision to train aLEXa in a multi-task setting.

Table 1: Binary classification results on designated test set.

	Precision	Recall	Macro F1 Score
Chalkidis et al. (2019)			
BERT	24.0%	50.0%	17.0%
HIER-BERT	90.4%	79.3%	82.0%
Haas and Skreta (2022)			
BERT	85.1%	**94.0%**	82.6%
RoBERTa	85.7%	93.9%	83.3%
LEGAL-BERT	86.3%	90.0%	81.8%
HIER-BERT (*1 layer*)	**91.3%**	80.4%	83.2%
HIER-BERT (*2 layer*)	**91.3%**	80.5%	83.3%
HIER-RoBERTa (*2 layer*)	89.9%	79.0%	81.7%
HIER-LEGAL-BERT (*2 layer*)	91.2%	80.5%	83.3%
aLEXa (*BERT base*)	91.1%	80.6%	**83.4%**
aLEXa (*RoBERTa base*)	83.9%	80.7%	81.9%

Discussion of quantitative results from multi-label classification. Meanwhile, for the multi-label classification task, we see similar results as o our baseline models outperform the models in Chalkidis et al. (2019), as displayed in Table 2. Specifically, our fine-tuned RoBERTa and LEGAL-BERT models outperform Chalkidis et al. and we achieve state-of-the-art results of with regard to the micro F1 (μ-F1) score of 62.1% using LEGAL-BERT. These results can again be explained due to Chalkidis et al. choices of learning rates. We also hypothesize that because the multi-label classification task is significantly more difficult than the binary prediction task, pre-training models from scratch on legal text corpora helps significantly, thus pushing LEGAL-BERTs performance.

The task difficulty is likely also the reason why our hierarchical models underperform the results set by Chalkidis et al. (2019) although achieving better results in the binary classification task. We expect that by adding more data or choosing fewer paragraphs in the hierarchical model, multi-label classification results can be further improved.

Table 2: Multi-label classification results on designated test set.

	Precision	Recall	μ-F1 Score
Chalkidis et al. (2019)			
HAN	65.0%	55.5%	59.9%
HIER-BERT	**65.9%**	55.1%	60.0%
Haas and Skreta (2022)			
BERT	63.9%	48.9%	55.4%
RoBERTa	63.5%	57.0%	60.1%
LEGAL-BERT	64.8%	**59.7%**	**62.1%**
HIER-BERT (*multi-head attn.*)	51.6%	47.5%	49.4%
HIER-RoBERTa (*2 layer*)	51.8%	56.0%	53.8%
aLEXa (*BERT base*)	56.1%	39.8%	46.5%

6   Analysis

In addition to the *quantitative* results outlined in Section 5.4, we devote substantial attention to a *qualitative* analysis of our results. By combining *quantitative* metrics with *qualitative* checks, we are able to increase our confidence in the quality of our work.

In their paper, Chalkidis et al. (2019) identify two key issues in legal judgement prediction: they state that most systems have severe limitations in "processing long documents" and provide "no

justification for their predictions" [2]. The former of these problem is solved by our approach: by building trainable hierarchical models which first embed paragraph meaning and then use multi-head attention or transformer layers to produce a final case embedding, we successfully process longer texts. In terms of the latter problem mentioned by Chalkidis et al. (2019), we also solved it with aLEXa. We go beyond paragraph attention to make legal fact selection an explicit component of the training procedure to improve the state of the art, which makes sense as justifications in the legal domain are most useful on a fact (paragraph) level as opposed to token-level attention scores.

To confirm that our model successfully justifies its predictions, we select a subset of precise examples, and analyze them "by hand" and by looking at attention over the paragraphs to better understand what the model gets wrong and whether there is any evidence of systematic bias in the model. We realize that aLEXa does indeed effectively select relevant paragraphs as visualized in Figure 5 below.

Case: Maxian And Maxianová v. Slovakia (2014).
Verdict: "Violation of Article 6 - Right to a fair trial (Article 6 - Civil proceedings; Article 6-1 - Reasonable time)"

4. The applicants are spouses. They were born in 1949 and 1965 respectively and live in Vienna, Austria.
5. On 13 April 2005 the applicants brought an action seeking dissolution of joint ownership of a real estate before the Dunajská Streda District Court (file no. 9C 70/2005).
6. On 6 September 2006, at its fifth hearing, the District Court delivered a judgment. The defendant appealed. The applicants requested the District Court to give a supplementary judgment. On 9 November 2006 the case file was submitted to the Trnava Regional Court.
7. On 20 March 2007 the Regional Court returned the case file to the first-instance court as incomplete. On 11 September 2007 the District Court gave a supplementary judgment and on 11 January 2008 the case file was again submitted to the Regional Court.
8. In 2008 the Regional Court stayed the proceedings for two months pending the outcome of inheritance proceedings after the defendant had died.
9. On 31 March 2009 the Regional Court quashed the first-instance judgment and remitted the case to the District Court for a new determination.
10. On 20 August 2010 the applicants complained before the Constitutional Court about the length of the proceedings before the District Court.
11. On 4 October 2010 the District Court approved the friendly settlement of the case reached between the parties. This decision became final on 30 October 2010.
12. On 24 November 2010 the Constitutional Court declared the applicants' complaint inadmissible as being manifestly ill-founded (case no. I. ÚS 455/2010). It held that there had been no significant delays in the proceedings before the District Court in breach of Article 6 § 1 of the Convention and its constitutional equivalent.

Figure 5: A sample visual verdict justification.

7 Conclusion

Main findings. Our state-of-the-art results for both the binary and multi-label classification tasks underscore the potential of domain pre-trained and hierarchical language models in legal judgement prediction. Our custom aLEXa model, a multi-task hierarchical language model with self-learning loss weights and attention forcing, teaches the model what legal facts to pay most attention to, underscoring the potential of explainability to redefine the state of neural LJP. Given limited time and computational resources available to us, we are confident we can further improve our results.

Limitations. Our work was mainly limited due to the constraints associated with our Azure virtual machine (VM). The training time made us limit the number of possible experiments we could conduct, while resource constraints limited us to using only about one third of the available words in the dataset for training. Additionally, we experienced longstanding VM connection issues, as well as site outages of Hugging Face that have temporarily suspended our work. Finally, multi-label hierarchical model performance remains a limitation, and its inefficiencies ought to be addressed in further research.

Future work. In addition to addressing the limitations outlined above, there are several potential avenues for future work in the field of neural judgment prediction. Our project could be neatly extended by anonymizing the data, similarly to the approach taken by Chalkidis et al. (2019) [2]. This could be accomplished by removing locations, names, and organizations using a named-entity recognition library (for instance, spaCy [13]), and thus helping understand and address the underlying model bias. Moreover, even though neural judgment prediction in English is a relative new era, there are many languages that have seen very limited use of legal NLP (oftentimes due to insufficient number and richness of the datasets available), which opens opportunities to popularize the field in many different jurisdictions. On a broader level, we are still in the very early innings of applying NLP models to specific expert domains, and advances on other domain-specific tasks could enable us to tangibly advance the mission of ameliorating societal problems through technology by creating the state of the art on tasks going beyond neural legal judgment prediction.

References

[1] Ramesh Nallapati and Christopher D. Manning. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

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

[3] Preoţiuc-Pietro D Lampos V Aletras N, Tsarapatsanis D. Predicting judicial decisions of the european court of human rights: a natural language processing perspective, 2016.

[4] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

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

[8] Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. A sentence-level hierarchical BERT model for document classification with limited labelled data. *CoRR*, abs/2106.06738, 2021.

[9] Qingyun Dou, Yiting Lu, Potsawee Manakul, Xixin Wu, and Mark J. F. Gales. Attention forcing for machine translation, 2021.

[10] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico, 2021. Association for Computational Linguistics.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115, 2017.

[12] Nikolaos Aletras and Ilias Chalkidis. Echr dataset, 2019.

[13] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.