From RoBERTa to aLEXa: The Gold Standard for Automated Legal Expert Arbitrators

Stanford CS224N Custom Project Milestone

Lukas Haas
Department of Computer Science
Stanford University
`lukhaas@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. We aim to establish better baselines for neural LJP by jointly fine-tuning large language models (LLMs) and training classification heads on top of them. Beyond these baselines, our goal is to improve the state of the art in neural LJP by building novel model architectures which employ attention forcing [1], teaching the model what legal facts to pay most attention to. We hope that this approach will also improve the explainability of LJP models, allowing us to both quantitatively and qualitatively assess the quality of legal predictions. For our milestone, we trained BERT, RoBERTa, and LEGAL-BERT models in a binary and multi-label classification setting. Thus far, our neural models outperform all binary and multi-label classification models used in the Chalkidis et al. (2019) paper [2]. We aspire to further push our results by the end of the project while contributing novel state-of-the-art findings to the LJP research community.

1 Key Information

- **TA Mentor**: Lucia Zheng.
- **External collaborators**: No.
- **External mentor**: No.
- **Sharing project**: No.

2 Approach

- **Goal**. For our milestone, our goal was to jointly fine-tune large language models (LLMs) and training classification heads on top of them for a variety of LLMs, with the goal of potentially outperforming some results in Chalkidis et al. (2019) [2].
- **Approach and baselines**. Following the feedback on our Project Proposal by our Project Mentor, Lucia Zheng, we used already pre-trained versions of **BERT** [3], **RoBERTa** [4], **LEGAL-BERT** [5] 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, which is a BERT model completely trained on legal text corpora. Although we experimented with using the RoBERTa large model from Hugging Face [4], we decided

against including that model in our approach as we ran into GPU memory problems even with extremely small batch sizes. The limited computational resources available to us on Microsoft Azure further buttressed our resolve to focus on the three former pre-trained models.

- **Downstream tasks**. We trained the above models on two human rights violation classification tasks, namely **binary classification** and **multi-label classification** in which the correct violated human rights article needed to be predicted out of the 21 different articles found in the training set. While the binary classification task was easy to conduct, we needed to build custom Hugging Face trainer classes to allow for multi-label classification.

3 Experiments

- **Data**: As outlined in our Project Proposal, we focus our work on the dataset presented in the Chalkidis et al. (2019) paper, i.e. the ECHR dataset which is publicly available [6] and split into pre-defined training, validation, and test subsets. The dataset contains approximately 11.5k cases with associated outcomes accessed from the European Court of Human Rights' public database. For each case, the dataset contains facts from the case description that were extracted using regular expressions. 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 of which 11 occur less than 50 times, and only 21 articles labels occurring in the training set. The distribution of the number of violations by article based on our exploratory data analysis are visualized in Figure 1, with Article 6 of the ECHR, the right to a fair trial, being predictably the most commonly seen one.
- **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 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), menaing 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 1.
- **Experimental details**: Due to BERT, RoBERTa, and LEGAL-BERT models being restricted to a maximum of 512 tokens, we fused all paragraphs for each legal case into a single text body which we then truncated by only using the first 512 tokens. As explained, we then finetuned the underlying model while simultaneously training the binary classification or multi-label prediction head on our downstream task. We experimented with different learning rates and found that a learning rate of $\alpha = 2 \cdot 10^{-5}$ seemed to give the best results.

  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. While not addressed by Chalkidis et al. in their paper, we hypothesize that the choice of learning rate in Chalkidis et al. (2019) [2] resulted in their BERT model underperforming even random baselines which can be seen in Table 1.
- **Results**: 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 models outperform all models in Chalkidis et al. (2019) when looking at the aggregate metric of macro F1, achieving a high score of 83.3% with RoBERTa.

  Meanwhile, for the multi-label classification task, we see similar results with respect to our models outperforming the models in Chalkidis et al. (2019), as displayed in Table 2. The LEGAL-BERT model outperforms with regard to the aggregate micro F1 (μ-F1) score of 62.1%. We 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.

Table 1: Binary classification results on designated test set.

	Precision	Recall	Macro F1 Score
Chalkidis et al. (2019) [2]			
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%

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

	Precision	Recall	μ-F1 Score
Chalkidis et al. (2019) [2]			
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%**

4 Future work

Although our work thus far has already shown great promise for redefining the state of the art in neural LJP, we plan to extend our current results by tackling a key constraint of all the three of the models we tried thus far. BERT, RoBERTa, and LEGAL-BERT are **all restricted to considering at most 512 tokens**. As suggested by our exploratory data analysis, the average paragraph contains approximately 50 words regardless of the article violation (as presented in Figure 2, and although the number of paragraphs per case varies considerably (as suggested by Figure 2), the number of words per case exceeds the model token restrictions by at least an order of magnitude on average. Consequently, we intend on creating a custom **hierarchical BERT model with attention forcing** [1]; the model will be able to input various paragraphs that could constitute the entire case in aggregate, and use attention mechanisms to address the relative importance of each paragraph for the case outcome.

Additionally, we plan to spend the remainder of our time understanding the **explainability** of our models by generating visual verdict justifications, which would permit us to grasp the models' rationales behind their decisions while enabling us to detect model biases. Finally, to remove bias from our models, we plan to follow the precedent set by Chalkidis et al. (2019) and anonymize our data by removing names, locations, etc by employing a named-entity recognition library such as spaCy [7]. Hopefully, this tri-partite extension based on a hierarchical BERT model with attention forcing, explainability and anonymization would enable us to tangibly advance the mission of improving judicial processes through technology by creating the state of the art in neural legal judgment prediction.

References

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

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

[4] 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.

[5] 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.

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

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

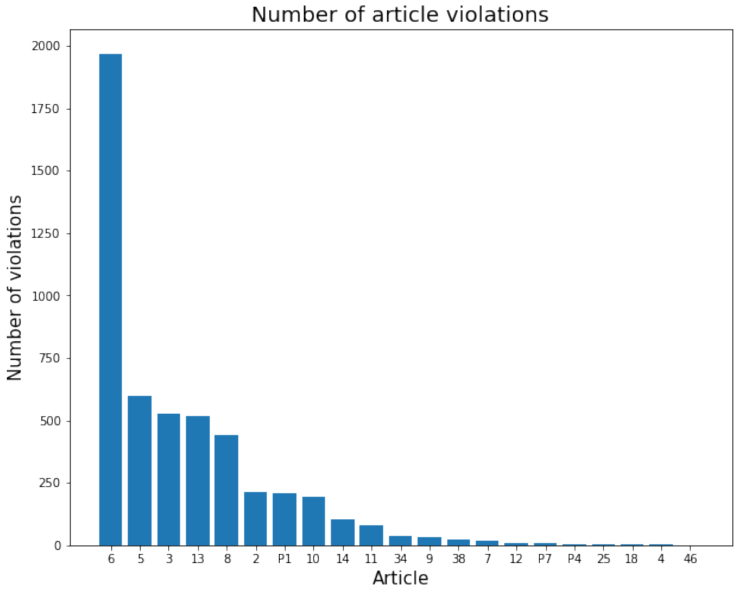

Figure 1: Number of article violations.

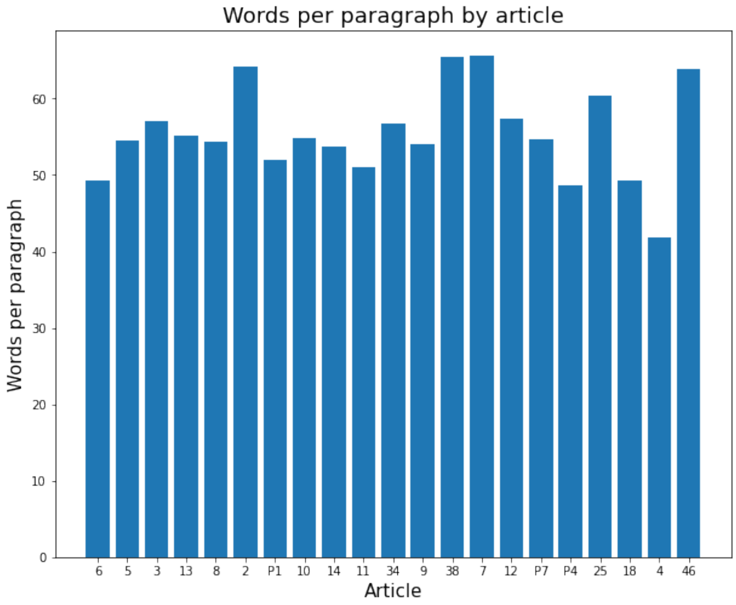

Figure 2: Words per paragraph by article.

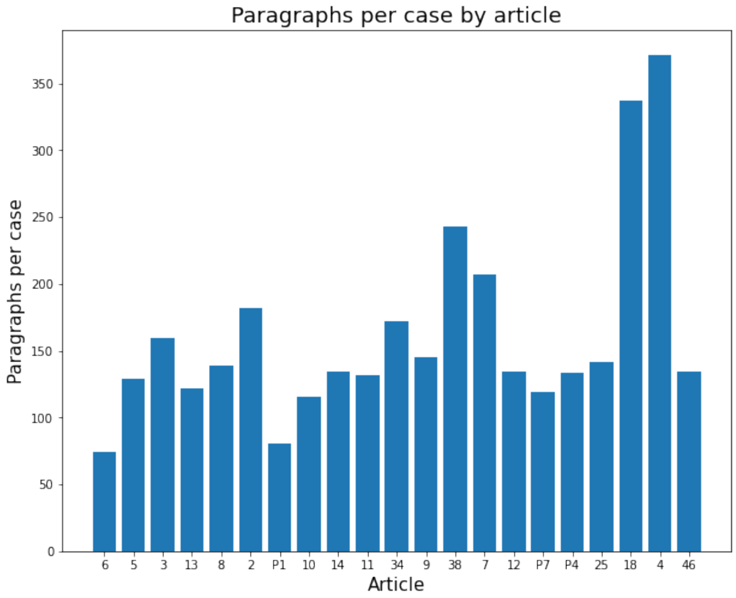

Figure 3: Paragraphs per case by article.