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

Stanford CS 224N Custom Project

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

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

1 Key Information

- *External collaborators (if you have any)*: No external collaborators.
- *Mentor (custom project only)*: (a) **Lucia Zheng** has already agreed to be our mentor.
- *Sharing project*: We are not sharing this project with any other course.

2 Research paper summary (max 2 pages)

Title	Neural Legal Judgment Prediction in English
Venue	Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
Year	2019
URL	https://aclanthology.org/P19-1424

Table 1: Bibliographical information of the Chalkidis et al. (2019) paper. [1].

Background. For as long as laws existed, they have been 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 [2]. 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. As Chalkidis et al. (2019) underscore, 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. The authors of the paper go beyond traditional, hand-crafted methods employed in legal NLP in English, and use end-to-end neural models to improve the state-of-the-art models in predicting the outcomes of legal cases.

Summary of contributions. Chalkidis et al. (2019) provide significant contributions to the field of legal judgment prediction. Previous papers 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 main contribution of the paper is the presentation of a novel dataset of 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.

Limitations and discussion. Despite their palpable contributions, Chalkidis et al. (2019)'s work nevertheless has some clear limitations. The key limitation lies in interpretability and explainability. Even though the neural models employed by the authors outperform the more traditional feature-based models, their predictions are black-box decisions without any justification. This presents us with an opportunity to focus specifically on understanding the rationale of the decisions made by our neural models. This could be done, for instance, by automating the analysis of other relevant legal material, and thus use multi-input systems to increase the robustness of the justifications behind the decisions. In addition to interpretability and explainability, the paper is limited in analogous ways to those in which other NLP papers are oftentimes constrained. This includes running the models on data in other languages than English as well as on data from other courts than ECHR. While the authors do not claim that their results are universally generalizable, the closer we get to the notion of generalizability in legal NLP, the more impact will our work be able to achieve in the real world.

Why this paper? We chose the *Neural Legal Judgment Prediction in English* paper by Chalkidis, Androutsopoulos, and Aletras due to its seminal nature in the field of legal judgment prediction. Before the paper was published in 2019, work on legal judgment prediction using neural models had been focused on Chinese or limited-size English datasets, significantly smaller than the novel dataset of 11.5k European Court of Human Rights (ECHR) cases provided by the authors. The paper was also available among papers on a list of legal NLP datasets suggested by our mentor, Lucia, which further buttressed our resolve to proceed with the paper. Finally, as we are both European and have an interest in applying deep learning to the legal space and social sciences, focusing on an application of NLP on a dataset from the European Court of Human Rights is particularly appealing.

Wider research context. Compared to many NLP papers that focus on flexible models that could be applied to a variety of tasks across domains, this paper is slightly different. It tackles a problem that is not domain-agnostic, but rather domain-specific. Consequently, it contributes to the broader literature on domain-specific training approaches. By showing the success of neural models in legal judgment prediction, it provides an incentive for other domains that rely on non-neural NLP models to revise their dominant approaches and to consider utilizing state-of-the-art language models fine-tuned to the specific domain to aid in the key prediction or classification tasks based on textual data in their fields. As we are still in the very early innings of applying NLP models to specific expert domains, it is an extremely exciting field that will benefit from humongous advances in the next few years, and we very much look forward to contributing to this dynamic field as outlined in the following section on our project description.

3 Project description (1-2 pages)

Goal. The goal of our project is to improve the state of the art in neural legal judgement prediction, answering the question of how different degrees of legal corpora pre-training and fine-tuning and prepossessing methods quantitatively and qualitatively affect the quality of legal predictions. As our project progresses and as time allows, our secondary goal is to explore adjacent areas of explainability, generating verdict justifications, or detecting model biases.

Our project's motivation is significantly influenced by the Chalkidis et al. (2019) paper which - for the first tine - showed promising results applying neural models to large legal text corpora in English. Before Chalkidis et al. (2019), such legal text corpora in English were largely unavailable, opening up an incredible opportunity for the computational analysis of legal texts.

Given that Chalkidis et al. employed BERT and hierarichal BERT [5] models trained on general English language corpora, we hypothesize that we can outperform their results using more powerful BERT variants such as RoBERTa [6] as well as employing BERT variants fine-tuned or completely trained on legal text data to generate a better model which we tentatively call the **A**utomated **L**egal **Ex**pert **A**rbitrator (aLEXa). By quantitatively and qualitatively comparing various degrees of model

domain-adaption and preprocessing methods in a binary and multi-class classification setting, we also hope to elucidate model decision making which is especially relevant in the legal domain.

Data. We will focus our work on the dataset presented in the Chalkidis et al. (2019) paper, i.e. the ECHR dataset which is publicly available [7]. 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.

The dataset provides a balanced train and development set and an unbalanced test set representing the real-world label distribution. Each case in the training set on average contains 43 paragraphs, 2,421 words, and 0.71 violated article labels.

Task. Our task for model development and evaluation is two-fold, consisting of a binary classification (1) and a multi-class classification (2) 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. Especially task (2) is challenging as the specific human rights articles which were violated do not occur with equal frequency in the dataset; 11 out of 66 article labels occur less than 50 times.

The following is an example of an input and output for a case which violated human rights. Most examples contain significantly more paragraphs. Empty `VIOLATED_ARTICLES` arrays mean that no ECHR article has been violated. The below example as been truncated to fit the page:

```
INPUT
["4. The applicant was born in 1951 and lives in Novi Pazar.",
"5. The applicant was employed by \u201cRa\u0161ka Holding Kompanija ...",
"6. On 22 December 2004, 5 January 2007, 10 April 2007, and 24 October ...",
"7. On 11 February 2005, 5 March 2007, 21 November 2007 and 31 December...",
"8. On 22 February 2005, 11 September 2007, 22 November 2007 and 3 ..."]

OUTPUT
VIOLATED_ARTICLES": ["6", "P1"]
```

Methods. Traditionally, applications of NLP in the legal space have struggled with two problems; first, legal texts are generally very long, containing dependencies stretching over many paragraphs which are difficult to model. Second, the legal language is very difficult to understand and contains special sentence structures and punctuation not commonly found in other English texts.

In order to address both of these problems, we aim to build on top of large, pre-trained foundational models such BERT with varying degrees of domain adaptation, comparing the performance of vanilla BERT models to models fine-tuned on legal corpora as well as BERT variants completely trained on legal texts such as Legal-BERT [8]. While we will be using pre-trained BERT and Legal-BERT models, we plan to fine-tuned BERT models to legal corpora ourselves.

Beyond the above methods, we aim to provide original contributions to the field of legal judgements predictions by experimenting with bespoke text preprocessing techniques targeted to legal texts and building attention mechanisms over the paragraph structure of cases, which will address the key limitation of the Chalkidis et al. (2019) paper regarding the automatic detection of critical paragraphs for the final court decision.

Beyond these methods, we plan to conduct a thorough quantitative and qualitative analysis of our results, elucidating our model's decisions, and discovering model biases by comparing the model's result on anonymized and un-anonymized versions of the ECHR corpora for which we will employ named-entity recognition (NER) frameworks in python.

Baselines. For both the binary and multi-class classification tasks, we plan to develop two specific baselines; first, for our simplest baseline, we will develop bag-of-words embeddings combined with SVMs and logistic regression models. The second baseline will be a vanilla implementation using BERT which we will build ourselves in order to compare our results to the stated BERT results in Chalkidis et al. (2019).

Evaluation. To evaluate our work, we plan to use both quantitative analysis as well as qualitative analysis to assess the validity of our results. For the quantitative part, we plan to follow the framework set out by the paper, and use F1 score evaluation for both binary violation classification and multi-class article violation classification tasks. In the multi-class setting, we will macro average the F1 scores for individual classes. By using the F1 scores, we would have a direct comparison with the baseline outlined above, and will be able to access our results vis-à-vis our cited paper. For the qualitative part, we plan to select a subset of precise examples, and analyze them "by hand" and by looking at attention over words to better understand what the model gets wrong and whether there is any evidence of systematic bias in the model. By combining quantitative metrics with qualitative checks, we would be able to more accurately evaluate the success of our work.

References

[1] 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.

[2] 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.

[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, 2019.

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

[8] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.