Automated Legal Expert Arbitrator for Neural Legal Judgment Prediction

Lukas Haas & Michal Skreta
{lukhaas, mskreta}@stanford.edu

Stanford ENGINEERING
Computer Science

CS 224N: Natural Language Processing with Deep Learning – Winter 2022

Problem Overview

- Although legal cases are usually represented in textual form, computational analysis has not been widely implemented in **legal judgment prediction**.
- Methods of **natural language processing (NLP)** based on **neural-network architectures** have shown impressive accuracy in predicting the outcomes of legal cases solely based on textual facts provided by the claimants [1].
- We build **transformer-based neural networks**, 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.

Background & Dataset

- Neural legal judgment prediction represents a relatively new field, with one of the first attempts in the area on **binary and multi-label classification problems** in English presented by **Chalkidis et al. (2019)** [1].
- We use a publicly available dataset from the **European Court of Human Rights (ECHR)** consisting of **11,478 cases** with associated outcomes as described in Chalkidis et al. [1]. For aLEXa (see Methods), we enrich this dataset with judgment rationales (relevant paragraphs) where available.
- In line with the paper, we opt for a pre-defined split of **7,100/1,380/2,998** cases between the **training**, **validation**, and **test** sets, respectively.
- For each case, the dataset contains a list of **paragraphs that constitute the case facts**, which have been extracted using regular expressions.
- Additionally, each case is mapped to violated articles of the **European Convention on Human Rights** with a total of **66** types of 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.

Experiments & Results

- We chose to evaluate our models using the ***macro* F1 score for the binary classification task** [Table 1] and the ***micro* F1 score metric for the multi-label task** [Table 2] in line with the original paper [1] and to address the multi-label class imbalance and to accurately compare results.
- **For both tasks, we achieve state-of-the-art results**, improving F1 scores by 1.3 and 2.1 percentage points for binary and multi-label, respectively.

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	91.2%	80.5%	**83.3%**

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%

Methods

- In approaching our problem, we trained models on **two downstream tasks** of human rights article violation – **binary (any article)** and **multi-label classification (specific article)** – in three increasingly complex steps:

1. We used pre-trained versions of three large language models from **Hugging Face**, i.e. **BERT**, **RoBERTa**, and **LEGAL-BERT** and fine-tuned them on our dataset, performing a hyperparameter grid search on data subsets. We only used the **first 512 tokens** of every case due to BERT-based token limits.
2. Next, we built **custom hierarchical models** using any of the above models as a base. Each **paragraph was fed through the base model** and the resulting embeddings were combined into a case embedding via **multi-head attention or transformer layers**, as per our specification in Fig. 1.
3. Finally, we introduced **Automated Legal Expert Arbitrator (aLEXa)**, a **multi-task hierarchical language model with self-learning loss weights** [2] using **attention forcing** [3] to learn legal judgement rationales (loss weighting function in Equation 1). aLEXa uses **BERT as the base model** and the Chalkidis 2021 dataset "rationales" for attention forcing, where available for each case [Fig. 2].

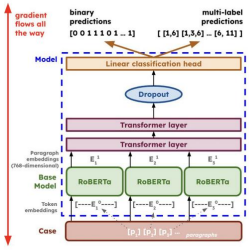

Figure 1: Hierarchical model architecture (base model can vary)

Equation 1: Loss weighting function in aLEXa based on [2].

$$\frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log\sigma_1\sigma_2$$

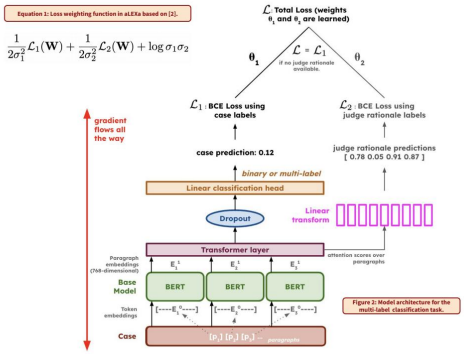

Figure 2: Model architecture for the multi-label classification task.

Analysis

- Two key issues in legal judgement prediction identified by Chalkidis et al. (2019) [1] are that most systems have severe limitations in "*processing long documents*" and provide "*no justification for their predictions*".
- 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**.
- **Justifications** in the legal domain are most useful on a **fact (paragraph) level** as opposed to token-level attention scores. By introducing **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**.
- Given our limited resources, through grid search on data subsets we found that processing **48 paragraphs with 224 word tokens** each using a **learning rate of 2e-5** worked best. This can likely still be improved.
- We also conducted a thorough qualitative analysis of **aLEXa**, showing that it **can effectively select the relevant paragraphs** in legal cases [Fig. 3].

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 3: Visualization of attention forcing over a non-training sample case.

Conclusions & Limitations

- 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.
- Given the **limited time and computational resources** available to us, we are confident we can further improve our results.
- Multi-label **hierarchical model performance** remains a limitation.

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] 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.

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