Argument & Computation 15 (2024) 113–119
DOI 10.3233/AAC-240002
IOS Press

Argument & Computation Community Resources (ACCR) Corner

Annotated insights into legal reasoning: A dataset of Article 6 ECHR cases

Jack Mumford *, Katie Atkinson and Trevor Bench-Capon
Department of Computer Science, University of Liverpool, UK

Abstract. We present a novel annotated dataset of legal cases pertaining to Article 6 – the right to a fair trial – of the European Convention on Human Rights (ECHR). This dataset will serve as a useful resource to the research community, to assist in the training and evaluation of AI systems designed to embody the legal reasoning involved in determining the appropriate legal outcome from a description of the case material. The annotations were applied to provide finer-grain classifications of legal cases at the document level. Each classification label was sourced from a domain knowledge model, derived with legal expert guidance and known as an Angelic Domain Model (ADM), such that the classifications correspond to the actual legal rationales used by the Court when determining the outcome of a given case. We discuss our annotation pipeline, including annotator training, inter-annotator reliability evaluation, and the dissemination of the annotation outputs and associated metadata.

Keywords: Legal case annotation, European Convention on Human Rights

1. Introduction

The European Convention on Human Rights (ECHR) has proved a popular domain for AI and Law research, especially for the use of machine learning applications (e.g. [2,7,13]). The domain, however, has also been addressed using more traditional symbolic techniques [10]. Although, of course, effort is required to build the model, the symbolic approach is able to achieve considerably higher accuracy. While the machine learning approaches rarely achieve more that 80% accuracy, the symbolic approach gives accuracy of over 90%. Moreover, symbolic approaches can readily explain their results, whereas explanations from machine learning approaches tend to be rather unsatisfactory [6]. This has led some to suggest a hybrid approach, using machine learning to ascribe factors, and a model relating these factors to the outcome to resolve the cases. In [6], the model is learnt by training a model on factors ascribed by domain experts, while in [15] and [16] a previously constructed symbolic model is used. The hybrid approach was first tried in [4], but there performance was not good: while the logical model achieved over 90% accuracy with manually ascribed factors, this fell to below 70% when factors were ascribed automatically. Since then, however, natural language techniques have improved immeasurably, and [6] claims good performance for this approach, albeit in a domain which uses fairly uniform terminology to describe cases and so is amenable to learning factor ascription.

*Corresponding author. E-mail: jack.mumford@liverpool.ac.uk.

The hybrid approach, however, requires annotated data with which to train the systems, and while the unannotated decisions of the ECHR are readily available,[1] annotated versions are not. To create such a dataset we undertook an exercise to annotate decisions with the elements of the symbolic model of Article 6 of the ECHR.

In the remainder of this paper, Section 2 provides some context around the approach taken to modelling Article 6 of the ECHR, Section 3 describes the resources themselves, and finally, Section 4 explains how the resources have been used as well as giving instructions on how to use them.

2. Context – modelling Article 6 of the ECHR

Although the ECHR has proved a popular domain for using machine learning approaches designed to predict the outcome of cases, such approaches have several problems:

- Explanation is, at best, rudimentary. Even when explanations are generated, as in [2], they lack a robust legal grounding and often hinge on spurious correlations such as dates and locations rather than on sound legal principles.
- Accuracy is not high: although some experiments have achieved results around 80%, the majority fall below this, often well below.
- The system may be using incorrect rationales [17], thus enforcing systematic wrong decisions.
- Performance tends to degrade over time [13]: the systems are trained on past decisions, and so cannot address the evolution of case law needed to predict future cases.

For these reasons, it can be argued that symbolic models retain an important place in such systems [14]. Legal reasoning within the ECtHR context, particularly under Article 6 – the right to a fair trial – demands a nuanced approach. Article 6 stands out not only due to its prominence as the article with the greatest number of applications, but also because of its procedural and objective nature, making it a fitting testbed for foundational work in developing practical, principled machine learning systems. There is a scarcity of that methodologies that can encapsulate expert-informed legal knowledge, an exception being the Angelic methodology ([1,5,10]), which captures the hierarchical reasoning structures used by legal experts to navigate complex cases, giving an Angelic Domain Model (ADM).

An ADM is crafted with the help of legal experts to reflect a factor-based legal model,as pioneered in [3], where *factors* are abstract patterns of fact that are used to establish prioritised sufficient conditions for the resolution of relevant legal issues for the case, which in turn provide necessary and sufficient conditions for the case outcome. The development of the ADM for Article 6 focuses on this structured reasoning, providing an explainable tool for determining case outcomes.

Recent research has established a hybrid architecture ([15,16]) that combines BERT-based NLP [11], specifically a hierarchical BERT structure (H-BERT) [12], with the ADM. The H-BERT NLP layer is used to interpret the case texts and ascribe findings to the leaf-factor nodes of the ADM, after which the internal logic of the ADM determines the case outcome. This hybrid structure not only allows for explainable predictions at a macro-level of legal abstraction via the ADM, but also affords micro-level insight through the attention mechanisms, which link case facts to specific ADM factors.

Preliminary findings suggest that this hybrid architecture enhances performance, especially in contexts with sparse data. Nonetheless, the success of the ADM factor ascription is highly dependent on the availability of robustly annotated datasets. Therefore, the need for meticulously annotated corpora cannot

[1] hudoc.echr.coe.int/

be overstated, as it provides the foundational knowledge enabling ML applications to bridge the gap between technical prediction and principled legal reasoning. In related published work, we describe the analysis of the annotation results (including inter-annotator scores) as well as the results of ML application to the annotated dataset [16].

3. Resource description

3.1. Overview

This repository presents the ADM for Article 6, together with a comprehensive dataset comprising annotations for cases related to this Article from the European Court of Human Rights (ECtHR). A total of 695 legal cases dating from 2015 onwards (in order to reduce the inclusion of out-of-date material) were annotated, 491 cases pertaining to violation of Article 6 verdicts and 204 pertaining to non-violation verdicts.[2] A unique aspect of this dataset is the meticulous annotation process conducted by 27 individual annotators, as reported in [16]. The annotators were final year undergraduate law students, 16 of whom had undertaken formal study on the ECHR (denoted as the domain group), with the remainder (denoted as the general group) not having take a course on this specific topic.[3]

The annotations delve into the intricate details of legal reasoning behind outcome of each case, providing a rich resource for analysis. The annotators demonstrated a high level of agreement, with a mean majority agreement score of over 95% (for more details see [16]). To facilitate research and application, the dataset includes summary annotations and a suite of scripts for both analysis and training a Hierarchical BERT (H-BERT) model [12]. This model aligns case descriptions with a legal knowledge model, the ADM, which explicates the reasoning behind court decisions.

3.2. Resource Content

(1) Individual Annotator Files:

- Comprises 54 sets of annotations, divided into two JSON files per annotator: one file for cases with a violation of Article 6 outcome and another file for non-violation outcomes.
- Each JSON file encapsulates a series of cases, enriched with detailed annotations presented as key-value pairs.

(2) Summary Annotation Files:

- Includes six JSON files offering a condensed version of the annotations: each JSON file corresponds to a specific combination resulting from two outcome types (violation and non-violation) and three different annotator group types (domain group, general group, and all annotators). The two outcome types are each paired with the three annotator groups, yielding a total of $2 \times 3 = 6$ unique files.
- These files amalgamate data from numerous cases, systematically organised as key-value pairs for ease of analysis.

[2]Note that any given case may or may not have resulted in a violation verdict of any other Article or Articles of the ECHR.

[3]Since the study involved human participants, we made a formal ethics application to, and were subsequently granted approval from, the University of Liverpool's Research Ethics Committee.

(3) Annotation Analysis Scripts:
- The repository contains three relevant Python scripts: one to assess annotator agreement, another to evaluate annotator productivity, and a third to analyze the distribution of case ascriptions to the ADM.

(4) H-BERT Training and Testing Scripts:
- Features three relevant Python scripts: one for normalising JSON annotation outputs, another for applying learning weights through the ADM, and a third for training and testing the H-BERT model on an Article 6 case corpus (corpus files available upon request).

(5) ADM Files:
- Contains two pivotal files: a PDF with a graphical representation of the ADM, used as a reference for annotators, and a CSV file detailing the ADM and its acceptance conditions, integral to the functioning of the provided Python scripts.

3.3. Novelty Compared to Existing Work

Comparative analysis with existing resources reveals that most annotations of ECtHR cases are structurally oriented, often employing regular expressions to categorise case text into a limited range of semantic segments. A notable resource in this domain is the work by Chalkidis et al. [9], which annotates cases based on silver and gold allegation rationales. However, these annotations do not follow a structured knowledge model to explain case outcomes, instead directly linking case facts to the outcomes.

In contrast, the resource we present in this paper is pioneering in its approach to annotating ECtHR cases. It uses our bespoke ADM, a domain-specific knowledge model, to elucidate the general reasoning employed in adjudicating Article 6 cases. This methodology not only offers a nuanced understanding of legal reasoning but also bridges the gap between factual analysis and outcome prediction in legal cases. In this current iteration of the dataset, the focus is on providing more fine-grained classifications at the document-level, via the ADM annotations, that will aid the development of tools for explainable and justifiable outcome classification and prediction. As such, our annotation exercise is very aligned with the Court's own efforts to establish key legal factors (which they denote as 'key words') that feature across the relevant case law.

A growing body of work has focused on predicting and classifying case outcomes from the facts using state-of-the-art NLP techniques, building on [13] and [8]. To our knowledge, ours is the first attempt to apply such a comprehensive and structured approach to the annotation of ECtHR cases, making it a unique and valuable contribution to the field of legal informatics and computational argumentation.

4. Resource use

4.1. Past Applications of the Resource

The primary application of this dataset has been in the training of Hierarchical BERT (H-BERT) Natural Language Processing models. This usage is prominently featured in the research documented in [16]. In this study, the dataset was the foundation for developing and refining NLP models capable of extracting rationales for case outcomes directly from the case material. The innovative approach employed in this research showcases the dataset's utility in enhancing the understanding and prediction of legal decisions based on the underlying reasoning.

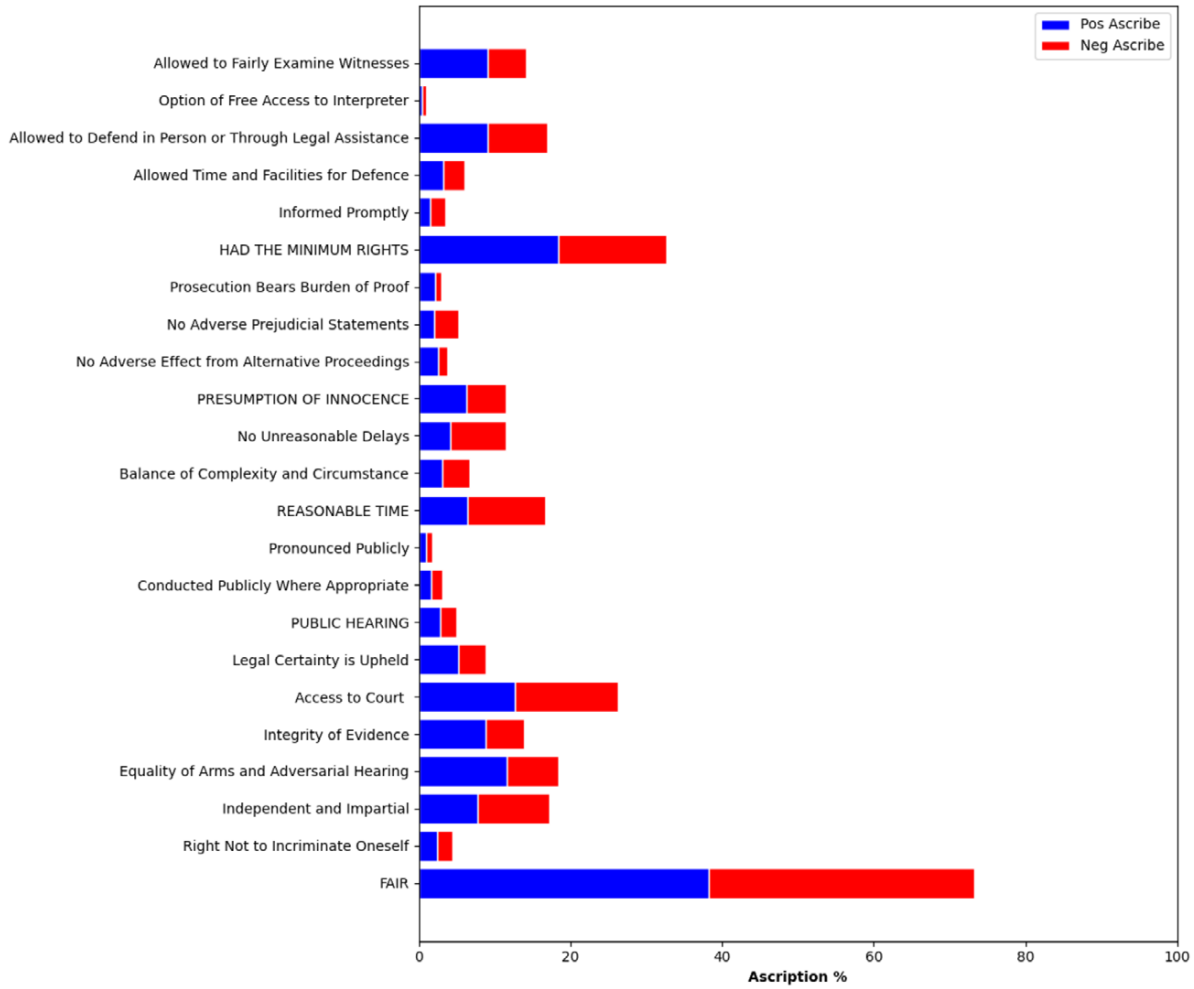

Fig. 1. Percentage of cases for which 'Issue' nodes (capitalised) and 'Intermediate' nodes (not capitalised) are ascribed. Any 'Intermediate' node pertains to the 'Issue' node most immediately positioned below it on the y-axis. 'Pos Ascribe' (resp. 'Neg Ascribe') is the percentage indicating the number of times the given node was annotated as corresponding to a non-violation (resp. violation) out of all non-violation annotations (resp violations).

The application of the dataset in this context demonstrates its effectiveness in contributing to advancements in legal informatics, particularly in the area of machine learning and NLP. By providing detailed annotations and a structured format, the dataset facilitates the development of models that can accurately interpret and analyse legal texts, offering significant insights into the legal reasoning process. For instance, the dataset provides novel insight into the distribution of legal factors raised as potential violations of Article 6. The dataset demonstrates significant variance between the distribution of factors, which researchers should understand and accommodate to avoid over-fitting to an unbalanced dataset that would be hidden if one only observes the outcome itself (see Fig. 1)

The ADM can also be used independently from the dataset, for example it has been used as a knowledge model in [18] to investigate the rationales of machine learning systems trained on incomplete and inconsistent data.

4.2. Availability and Instructions for Prospective Users

The dataset, along with its associated scripts and tools, is readily available for use and can be accessed through our dedicated GitHub repository.[4] This repository is designed to be user-friendly and includes comprehensive instructions for installation and usage, ensuring that researchers and practitioners can effectively use the resource for their specific needs. To access and use the dataset:

(1) Accessing the Repository:

- Visit the GitHub repository at https://github.com/jamumford/ECHR_Article6_ADM_Ascribe.
- The repository contains all necessary files, including the dataset, scripts, and ADM files.

(2) Installation:

- Follow the step-by-step installation guide provided in the repository to set up the environment and dependencies required for using the dataset and scripts.

(3) Usage:

- The repository includes detailed instructions on how to use the dataset for various applications, including training H-BERT models and analysing annotations.
- Users can refer to the provided examples and documentation to understand how to integrate the dataset into their research or application.

(4) Support and Updates:

- For additional support or queries, users can reach out through the contact information provided in the repository.
- The repository is actively maintained, with regular updates and enhancements to ensure its ongoing utility and relevance to the field.

Acknowledgements

We are grateful to Kanstantsin Dzehtsiarou and Jeremy Marshall from the School of Law and Social Justice at the University of Liverpool for their assistance in setting up the study on which the annotated dataset reported in this paper was produced. This work was supported by Towards Turing 2.0 under the EPSRC Grant D-ACAD-052 & The Alan Turing Institute.

References

[1] L. Al-Abdulkarim, K. Atkinson and T. Bench-Capon, A methodology for designing systems to reason with legal cases using ADFs, *AI and Law* **24**(1) (2016), 1–49.
[2] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro and V. Lampos, Predicting judicial decisions of the ECHR: A natural language processing perspective, *PeerJ Computer Science* **2** (2016), e93. doi:10.7717/peerj-cs.93.
[3] V. Aleven, Teaching case-based argumentation through a model and examples, Ph.D. thesis, University of Pittsburgh, 1997.
[4] K.D. Ashley and S. Brüninghaus, Automatically classifying case texts and predicting outcomes, *AI and Law* **17**(2) (2009), 125–165.

[4] https://github.com/jamumford/ECHR_Article6_ADM_Ascribe

[5] K. Atkinson and T. Bench-Capon, ANGELIC II: An improved methodology for representing legal domain knowledge, in: *Proceedings of the 19th ICAIL*, 2023, pp. 12–21.
[6] L.K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff and B. Liao, Scalable and explainable legal prediction, *AI and Law* **29**(2) (2021), 213–238.
[7] I. Chalkidis, I. Androutsopoulos and N. Aletras, Neural legal judgment prediction in English, 2019, arXiv preprint arXiv:1906.02059.
[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras and I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, 2020, arXiv preprint arXiv:2010.02559.
[9] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos and P. Malakasiotis, Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 226–241.
[10] J. Collenette, K. Atkinson and T. Bench-Capon, Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights, *Artificial Intelligence* **317** (2023), 103861. doi:10.1016/j.artint.2023.103861.
[11] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[12] J. Lu, M. Henchion, I. Bacher and B.M. Namee, A sentence-level hierarchical BERT model for document classification with limited labelled data, in: *Proceedings of DS 2021*, Springer, 2021, pp. 231–241. doi:10.1145/3478905.3478951.
[13] M. Medvedeva, M. Vols and M. Wieling, Using machine learning to predict decisions of the European Court of Human Rights, *AI and Law* (2019), 1–30.
[14] J. Mumford, K. Atkinson and T. Bench-Capon, Machine learning and legal argument, in: *CEUR Workshop Proceedings*, Vol. 2937, 2021, pp. 47–56.
[15] J. Mumford, K. Atkinson and T. Bench-Capon, Reasoning with legal cases: A hybrid ADF-ML approach, in: *Proceedings of JURIX 2022*, 2022, pp. 93–102.
[16] J. Mumford, K. Atkinson and T. Bench-Capon, Combining a legal knowledge model with machine learning for reasoning with legal cases, in: *Proceedings of the 19th ICAIL*, 2023, pp. 167–176.
[17] C. Steging, S. Renooij and B. Verheij, Discovering the rationale of decisions: Towards a method for aligning learning and reasoning, in: *Proceedings of the 18th ICAIL*, ACM, 2021, pp. 235–239.
[18] C. Steging, S. Renooij and B. Verheij, Improving rationales with small, inconsistent and incomplete data, in: *Proceedings of Jurix 2023*, 2023, pp. 53–62.