Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases

Jack Mumford
University of Liverpool
Liverpool, UK

Katie Atkinson
University of Liverpool
Liverpool, UK

Trevor Bench-Capon
University of Liverpool
Liverpool, UK

ABSTRACT

Recent years have witnessed significant progress in the deployment of advanced Natural Language Processing (NLP) techniques based on transformer technology, across many domains and applications. However, in legal domains, due to the complexity, length, and sparsity of legal case documents, the use of these advanced NLP techniques has offered comparatively slight returns. Perhaps even more importantly, such methods are critically lacking in explainability and justification of outputs, which are essential for many legal applications. We propose that the direction of these NLP techniques should be aimed at ascription to a legal knowledge model, which can then provide the necessary and auditable justifications for the rationale of any case outcome. In this paper we investigate the effectiveness of using Hierarchical Bidirectional Encoder Representations from Transformers (H-BERT) models to ascribe to an Angelic Domain Model (ADM) that is able to represent the legal knowledge of a domain in a structured way, enabling justifications and improving performance. Our study involved an annotation task on a popular domain, cases from the European Court of Human Rights, to gain an understanding of the balance of complaints in the domain. The data set produced from this study enabled training of models for factor ascription using the classification targets derived from the annotations. We present results of experiments conducted to evaluate the performance of the ascription task at three different levels of abstraction within the structured model.

CCS CONCEPTS

• **Applied computing → Law**.

KEYWORDS

Transformers, domain model, factor ascription, case annotation

ACM Reference Format:
Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2023. Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595158

1 INTRODUCTION

The ECtHR (European Court of Human Rights) has featured in a number of studies within the AI and Law literature with the goal of developing AI systems that are able to provide decision support for reasoning about and deciding legal cases. Some studies, e.g. [4] and [20], have tackled this task as a classification exercise using machine learning techniques to determine which cases are violations of a particular article of the convention and which are non-violations. Other work, e.g. [14], has shown how to build a symbolic model of the domain and use this within an account of factor-based reasoning [12] to decide cases within the domain. Due to the potential complexity of legal cases (ECtHR cases are frequently multiple pages in summary form), a symbolic model must focus on an appropriate level of high-level abstraction in order to capture the relevant reasoning in a structured manner, without becoming too unwieldy for human-friendly explanation or too brittle for successful use.

In this paper we investigate the task of ascription of legal cases by combining a symbolic domain model, that we call an ADM (Angelic Design Model), with a H-BERT (Hierarchical Bidirectional Encoder Representations from Transformers) model[19] that uses state-of-the-art natural language processing techniques. The aim of the work is to use these techniques with the legal knowledge of a domain represented in a structured way to enable justifications and improve performance of the automation of reasoning about legal cases. The symbolic ADM is able to justify the outcome of a case by connecting it to the nodes of the ADM, where the nodes are the key legal issues and factors representing stereotypical patterns of facts relevant to the particular legal domain. An ADM would be too unwieldy and brittle if it were designed to encompass all the potential facts relevant to a particular legal domain. Hence, we target ML (machine learning) at the task of ascribing legal factors and issues from the facts, but also require this application of ML to be explainable and grounded in law [22]. We select H-BERT models for the ML application, due to their potential ability to highlight relevant passages of text to justify their ascription output.

Specifically, we leverage H-BERT models to process the natural language descriptions of the facts of ECtHR cases and output the key legal factors and issues that justify the case outcome, according to the ADM. We evaluate the H-BERT models against ascription classification targets derived from an annotated data set produced by law students tasked with labelling ECtHR cases with the relevant factors and issues from the ADM. We produce encouraging results indicating the ability of the H-BERT models to effectively ascribe to the ADM, with the strongest performance attained for those legal factors and issues more frequently attributed to cases in the annotations.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06.
https://doi.org/10.1145/3594536.3595158

In section 2 we provide an overview of the context of our research, including the prior work that we make use of. In section 3 we provide a detailed description of an annotation exercise we have undertaken to produce an annotated data set of a popular domain of study in AI and Law, the ECHR (European Convention on Human Rights). The purpose of the study is to label ECHR cases submitted under Article 6 (the right to a fair trial), in accordance with an existing structured model of the domain. We describe the study and its outcomes, key of which is a data set that provides classification targets for an AI system and also reveals the distribution of legal factors and issues across the corpus used. In section 4 we describe an implementation that has been enabled by the annotated data set for training H-BERT models for factor ascription using the classification targets derived from the annotations. In section 5 we evaluate the performance of the H-BERT models in the ascription task at three different levels of abstraction within the ADM. Section 6 provides a discussion of the results of the experimental evaluation and section 7 closes the paper with some concluding remarks and steps for future work.

2 BACKGROUND

The setting for the work reported in this paper is the body of literature in AI and Law on reasoning about legal cases. This topic is a long standing one within the field and has evolved as techniques for representing and reasoning about legal cases have developed and matured. The starting point for the proposals that we set out here is the desire to ensure that the AI-based tools we build are capturing legal reasoning processes. To achieve this, similar to many prominent works prior to ours (e.g. HYPO [27], CATO [5] and IBP [12]), we use of symbolic AI techniques to capture domain knowledge in our models, to support procedures that enact reasoning about the features of legal cases. Use of such techniques has been shown to ensure that reasoning processes employed in knowledge-based systems can yield explanations of outcomes that are readily understood by end users, a topic discussed in detail in [6].

However, in recent years there has been an increasing number of papers that are focussed on the legal case prediction problem, see e.g. [21] and [30]. In this approach, machine learning algorithms are trained on data sets to classify cases based on similarities between data points that represent the cases. This approach does not encode legal reasoning in the same manner as symbolic techniques but has nonetheless become popular as large data sets for the training exercise, such as Article 6 of the ECHR, have become more readily available. Rapid recent advances within the field of Natural Language Processing assist with the identification of datapoints from source texts.

Given the maturation of these two strands of research, symbolic and subsymbolic, aimed at providing automated assistance for deciding legal cases, the work we present in this paper is a step towards harnessing the benefits of the two approaches and combining them into a hybrid system. Before demonstrating how we approach this task, we first provide a short overview of the foundational work that we use as our starting point.

2.1 Angelic Design

The Angelic methodology [2] was developed to enable structured modelling of the reasoning within a legal domain. The inspiration for the methodology came from the Abstract Dialectical Frameworks (ADFs) of Brewka and Woltran ([11] and revised in [10]).

ADFs are a generalisation of Dung's abstract argumentation frameworks [17]. ADFs comprise a three tuple: a set of nodes, a set of directed links joining pairs of nodes (a parent node and its child nodes), and a set of acceptance conditions, expressed in terms of the children. The links show which nodes are used to determine the acceptability (or otherwise) of any particular node, so that the acceptability of a parent node is determined solely by its children.

The key idea is that by adding acceptance conditions to the nodes of an abstract factor hierarchy, of the sort developed in CATO [5], we can obtain an ADF. Thus we are able to represent the structure of a legal domain using the nodes and edges to capture the knowledge expressed by the factor hierarchy, with the acceptance conditions able to capture knowledge of how the children relate to their parents, derived from statutes and cases. Thus, used in the legal context, the nodes represent statements which relate to the issues, intermediate factors and base level factors as expressed in CATO's factor hierarchies. For determining the outcome from the issues, the acceptance conditions are standard expressions of propositional logic, so that the upper levels of the structure form a logical model as found in IBP [12]. It was shown in [26] that precedents can be represented as a set of prioritised rules. For resolving issues in terms of factors, such rules can form the basis for the acceptance conditions, with the priorities being determined by the decisions made in precedent cases. For leaf nodes in the ADF, acceptance and rejection is determined by the user, on the basis of the facts of the particular legal case that is under consideration. Collectively, the acceptance conditions can been seen as a knowledge base, but the ADF provides a high degree of modularity, since each node contains all and only the information needed to determine its acceptance or rejection. This is important to enable the domain knowledge captured in an ADF to be easily modified and updated as the law evolves over time [1]. Additionally, the acceptance conditions are used to generate arguments about the case and the ADF structure guides the deployment of the arguments; these arguments can form the basis of an explanation of the outcome, both verbally as in [2] and [14], and graphically as in [8].

It has been shown that ADFs can been applied in law to model factor-based reasoning in a number of domains popular in the academic literature [2], real world applications provided by a law firm [3] and most recently, cases from the European Court of Human Rights (ECtHR) [14]. In applying the methodology within these studies, Angelic has been extended to accommodate additional information in the nodes (e.g. the sources of the various acceptance conditions as in [7]) and to include a list of questions which elicit facts from the user that can be used to determine whether or not the base level factors are satisfied, as, for example, in [8]. This enables the user to be guided with information as to how factors should be ascribed. As a result of the extensions, and because the form of the ADF is restricted, we now refer to the structure as an *Angelic Domain Model* (ADM), rather than as an ADF.

Table 1: Number of cases annotated per hour during the research study. The returns of students in the group with domain knowledge of the ECHR are indicated by 'Domain', and those from the group without ECHR domain knowledge are indicated by 'Non-domain'. Mean and standard deviation results are reported to 2 d.p., whereas results are reported to 3 d.p. for t-test p values. The far column t-test reports the one-sided p value associated with the null hypothesis that the mean value for cases per hour is equal or higher for violation cases than non-violation cases. The two bottom t-tests (vio and non respectively) report the one-sided p value associated with null hypothesis that the Domain mean (for violation and non-violation cases respectively) is equal or greater than the Non-domain mean.

	Violation	Non-violation	t-test
Domain cases/hour	5.62	8.60	0.005
Domain stdev	2.38	3.62	
Non-domain cases/hour	6.61	10.15	0.015
Non-domain stdev	2.88	4.09	
t-test (vio)	p value = 0.169		
t-test (non)	p value = 0.155		

3 PRODUCING AN ANNOTATED DATA SET

In order to effectively ascribe legal factors and issues to a corpus of legal cases, we need to be able to verify that any ascription method is performing appropriately. To that end, we performed a study in which we recruited 27 final-year law undergraduate students drawn from two groups, one with domain knowledge and the other without. The students were tasked with reading the corpus of cases pertaining to Article 6 of the ECHR – concerning the right to a fair trial – and labelling the cases in accordance with the ADM. Since the ADM was drafted from expert knowledge and documentation deemed accurate from January 2015 onwards (see Section 4 and Tables 4-8 for an outline of the ADM), we restricted the corpus to cases concluded after this date, resulting in 530 violation cases and 205 non-violation cases extracted from HUDOC[1] – the principal repository for ECHR legal case documentation. Thus, we were able to derive a data set that both provides classification targets for an AI system and also reveals the distribution of factors and issues across the corpus, which can be used to gain an understanding of the balance of complaints in the domain.

We had two objectives when recruiting the students for the annotation research study:

(1) To produce a reliable annotated data set with maximal coverage.
(2) To investigate the impact of greater domain knowledge on the ability to annotate the corpus effectively.

To satisfy the first objective, we sought to balance workload per student, to ensure students were annotating a sufficient number of cases for their performance to be reasonably evaluated, and the total number of students, to similarly ensure a sufficient sample size for evaluation. We therefore recruited the 27 students to work two hours each evening for a full working week (with the first evening consisting of the training). To improve the productivity of the students, they were explicitly instructed not to read the entire text for any given case, but to only read THE LAW section (see Figure 1 for an example excerpt) that describes the court's reasoning of the case and thus is aligned with the reasoning contained in the ADM. In terms of ascription, three labels are available for a node in the ADM for any given case:

- positive ascription, an ADM node is relevant and satisfied.
- negative ascription, an ADM node is relevant but not satisfied (which for our model will cause a violation).
- non-ascription, an ADM node is not relevant.

To save time, students were only required to label for positive and negative ascription: non-ascription was taken as implicit from an empty label. We also ran a pilot prior to the main study, which revealed that annotating at the leaf factor level necessitated a high cognitive load that would severely reduce the productivity of the students. The decision was therefore taken to annotate intermediate factors as the lowest abstraction level as a sensible compromise between productivity and granularity. Prior to the commencement of the research study, we had expected the students to be able to process approximately three cases per hour, but the students were able to achieve a significantly higher productivity, as indicated in Table 1. Accordingly, over the course of the research project, students were able to provide annotations for 204 of the 205 non-violation cases, and 491 out of the 530 violation cases, providing substantial coverage beyond our initial expectations.

In terms of the reliability of the annotations, we must first consider the scope of the annotation task. Students were provided with training with the use of the ADM and instructions for annotating at an intermediate factor abstraction level, one level higher than the base-level leaf factors. We selected this abstraction level based on a pilot study that suggested it as the optimal point for balancing comprehension on the annotator's part and meaningful granularity in terms of explanation power. In comparison, abstraction at the issue level offered even easier comprehension with only five nodes to understand, but was too high-level from an explainability

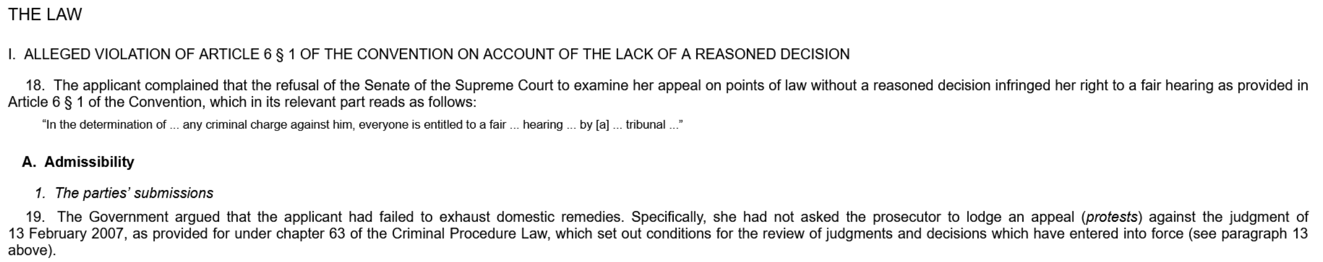
Table 2: Inter-annotator agreement Fleiss' kappa score across violation and non-violation cases (results reported to 3 d.p.).

	kappa score
Violation	0.551
Non-violation	0.570

Table 3: Majority agreement score across the two groups. The first two rows present the means and standard deviations (results reported to 2 d.p.). The t-test reports the one-sided p value associated with the null hypothesis that the Non-domain mean is equal or greater than the Domain mean (result reported to 3 d.p.).

	Mean	Stdev
Domain	95.61	1.02
Non-domain	95.09	1.23
t-test	p value = 0.123	

[1] hudoc.echr.coe.int/



THE LAW

I. ALLEGED VIOLATION OF ARTICLE 6 § 1 OF THE CONVENTION ON ACCOUNT OF THE LACK OF A REASONED DECISION

18. The applicant complained that the refusal of the Senate of the Supreme Court to examine her appeal on points of law without a reasoned decision infringed her right to a fair hearing as provided in Article 6 § 1 of the Convention, which in its relevant part reads as follows:

> "In the determination of ... any criminal charge against him, everyone is entitled to a fair ... hearing ... by [a] ... tribunal ..."

A. Admissibility

1. The parties' submissions

19. The Government argued that the applicant had failed to exhaust domestic remedies. Specifically, she had not asked the prosecutor to lodge an appeal (*protests*) against the judgment of 13 February 2007, as provided for under chapter 63 of the Criminal Procedure Law, which set out conditions for the review of judgments and decisions which have entered into force (see paragraph 13 above).

Figure 1: Excerpt from THE LAW section of TALMANE v. LATVIA, an ECHR Article 6 case annotated in the research study.

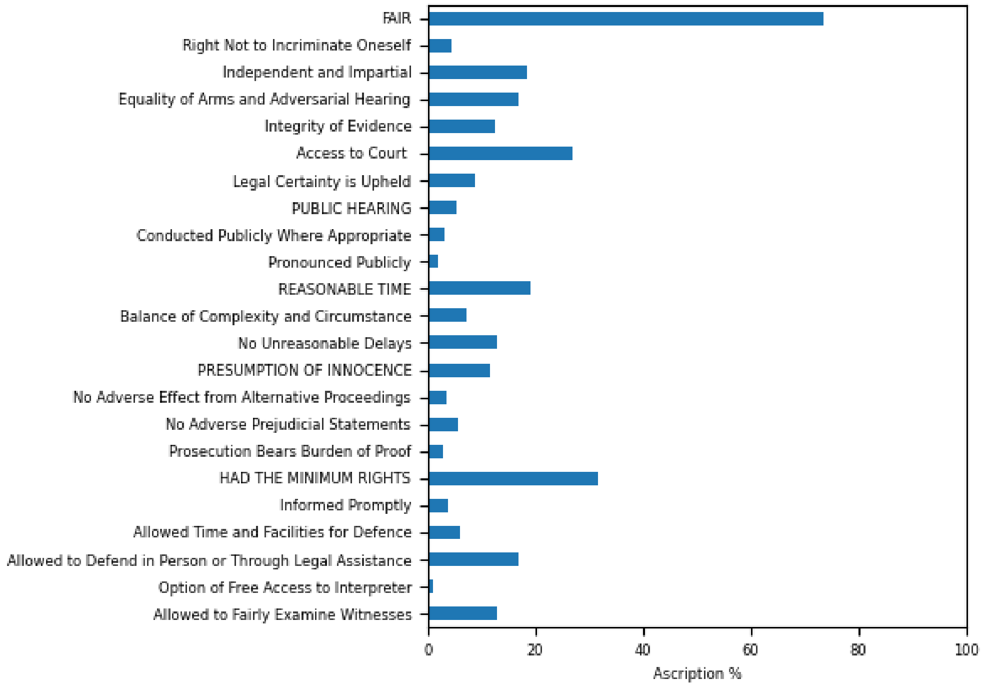


Figure 2: Percentage of cases for which 'Issue' nodes (capitalised) and 'Intermediate' nodes (not capitalised) are ascribed. Any 'Intermediate' node pertains to the 'Issue' node most immediately positioned above it on the y-axis.

perspective. And the leaf factor abstraction level offers the greatest explainability but requires the annotator to bear a significant cognitive load of continuous recall of not only the 5 issue nodes and the 18 intermediate factor nodes, but also the 34 leaf factor nodes, in order to make informed annotations. Of course, since the ADM offers a hierarchy of reasoning, by annotating at the intermediate factor level students were also simultaneously producing annotations at the issue level.

To measure and evaluate the reliability of the students' annotations we use two metrics, namely: Fleiss' kappa; and a majority agreement score. Fleiss' kappa was selected as suitable for our study since it is a widely respected measure for inter-annotator agreement when assessing the reliability of agreement between a fixed number of annotators, for a fixed number of items, which do not require annotation by every annotator. Table 2 presents the Fleiss' kappa scores for the violation and non-violation outcome types of cases in the corpus. The scores are very similar, indicating similarly strong performance across the corpus regardless of outcome. Whilst there is no definitive consensus of how to interpret the Fleiss' kappa (beyond 0 indicating no agreement and 1 indicating perfect agreement), multiple sources cite an interpretation advocated in [29], which if used for our scores would indicate moderate bordering on substantial agreement. However, this interpretation is highly subjective and restricted to application to data sets consisting of only a few items. As our students were annotating a data set consisting of 23 items (5 issues and 18 intermediate factors) we consider the Fleiss' kappa scores in Table 2 to be much stronger than the interpretation of [29] would suggest, indicating substantial levels of agreement. The Fleiss' kappa scores also set a benchmark against which similarly scaled annotations can be assessed. The majority agreement scores presented in Table 3 provide another indication of the reliability of the annotations. The majority agreement scores switch the focus onto the reliability of the student annotator rather than on the individual items being annotated. The scores are calculated as the percentage of annotations made by the annotator that were the same as the majority consensus across all other annotators (discounting unanimous implicit annotation of non-ascription items). Table 3 clearly shows very high average agreement scores with relatively small standard deviation, indicating consistent majority consensus over the corpus, which encourages high confidence that the data set can be considered reliably annotated in accordance with the ADM.

To satisfy the second objective, we recruited final-year law students and divided them into two groups based on whether they had studied a module specifically focused on the ECHR. Out of the 27 students, 16 had undertaken the module and formed the domain group, with the other 11 students forming the non-domain group. As previously discussed, we measured the students' productivity (cases annotated per hour) and the reliability of their annotations. In terms of productivity, Table 1 indicates that, perhaps counter-intuitively, the non-domain group were on average more productive. This pattern was evident across both violation and non-violation cases, with similar t-test p values that suggest a reasonable degree of confidence that the relationship of higher productivity for the non-domain group is statistically significant. Formally, we reject the null hypothesis that the mean of the domain knowledge population is higher or equal to that of the non-domain knowledge population, at least at the 80% confidence level.

In terms of the reliability of the annotations between the two groups, Table 3 suggests an intuitive result that the group with domain knowledge were generally more reliable in their annotations. Both groups achieved high majority agreement scores, but the t-test p value suggests that we can reject the null hypothesis that the mean of the non-domain knowledge population is higher or equal to that of the domain knowledge population, at least at the 85% confidence level. This suggests an even stronger statistical significance for the observation that the domain group were more reliable in their annotations, than we observed for the higher productivity of the non-domain group.

Finally, Figure 2 displays the distribution of issues and intermediate factors across the corpus as annotated by the research study cohort. Whilst the students distinguished between positive and negative ascription when labelling, we have merged these two ascription labels to align with the focus on ascription vs non-ascription relevant to this paper. It is clear that there is significant variance of proclivity for issues and factors raised as complaints under Article 6. The distribution captured in Figure 2 could be useful for identifying trends for groups interested in the dynamics of the ECHR and could be further examined alongside other parameters such as geography of complaint. It could also prove useful in providing estimated annotations for cases, supporting unsupervised/semi-supervised learning for explainable case classification with reference to the reasoning expressed in the ADM.

Table 4: Depiction of 'Intermediate' nodes relating to the 'FAIR' issue of the ADM. The 'Intermediate' nodes are listed with their acceptance condition (AND, OR) with respect to satisfaction of their children. The children are in turn the 'Leaf' nodes of the ADM.

Right Not to Incriminate Oneself (AND)
• Not Compelled to Testify • No Subterfuge • Testimony with Knowledge of Rights
Independent and Impartial (AND)
• Functional Nature • Personal Nature
Equality of Arms and Adversarial Hearing (AND)
• Fair Balance in Presenting Case • Access and Comment on Evidence
Integrity of Evidence (AND)
• Evidence Fairly Obtained • No Reasonable Concerns for Other Articles • Principle of Immediacy is Upheld
Access to Court (OR)
• Opportunity for Tribunal • Legitimate Reasons for Limitations
Legal Certainty is Upheld (AND)
• Legally Binding Where Appropriate • No Conflicting Decisions

Table 5: Depiction of 'Intermediate' nodes relating to the 'PUBLIC HEARING' issue of the ADM.

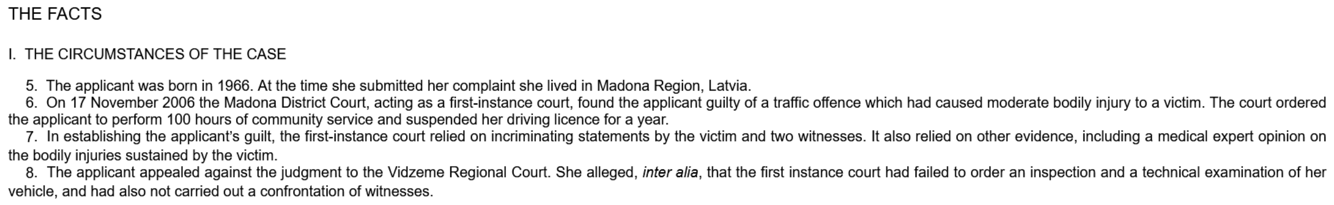
Conducted Publicly Where Appropriate (OR)
• Not Conducted Publicly And Reasonable Concern • Conducted Publicy And No Reasonable Concern
Pronounced Publicly (AND)
• Commensurate with Any Concerns

Table 6: Depiction of 'Intermediate' nodes relating to the 'REASONABLE TIME' issue of the ADM.

Balance of Complexity and Circumstance (AND)
• Complexity of the Case • Commensurate with Stakes
No Unreasonable Delays (OR)
• Delays Responsibility of Applicant • Delays Justified

4 IMPLEMENTATION

With the annotated data set produced through the research study described in the previous section, we are able to train H-BERT models for factor ascription using the classification targets derived from the annotations. However, unlike the annotation task (where annotators read solely from THE LAW section of a case), the H-BERT models are provided with text input solely from THE FACTS section (see Figure 3 for an example excerpt) of any given case. This is because ascription of factors and issues progresses through



THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in 1966. At the time she submitted her complaint she lived in Madona Region, Latvia.

6. On 17 November 2006 the Madona District Court, acting as a first-instance court, found the applicant guilty of a traffic offence which had caused moderate bodily injury to a victim. The court ordered the applicant to perform 100 hours of community service and suspended her driving licence for a year.

7. In establishing the applicant's guilt, the first-instance court relied on incriminating statements by the victim and two witnesses. It also relied on other evidence, including a medical expert opinion on the bodily injuries sustained by the victim.

8. The applicant appealed against the judgment to the Vidzeme Regional Court. She alleged, *inter alia*, that the first instance court had failed to order an inspection and a technical examination of her vehicle, and had also not carried out a confrontation of witnesses.

Figure 3: Excerpt from THE FACTS section of TALMANE v. LATVIA, an ECHR Article 6 case annotated in the research study.

Table 7: Depiction of 'Intermediate' nodes relating to the 'PRESUMPTION OF INNOCENCE' issue of the ADM.

No Adverse Effect from Alternative Proceedings (AND)
• Parallel Or Previous Do Not Jeopardise
No Adverse Prejudicial Statements (AND)
• Officials Do Not Undermine
Prosecution Bears Burden of Proof (OR)
• Beyond Reasonable Doubt • Criminal Liability Justified • Civil Liability Justified

Table 8: Depiction of 'Intermediate' nodes relating to the 'HAD THE MINIMUM RIGHTS' issue of the ADM.

Informed Promptly (AND)
• Informed of Accusation • Informed of Details Circumstances
Allowed Time and Facilities for Defence (AND)
• Adequate Time and Facilities to Organise Defence
Allowed to Defend in Person or Through Legal Assistance (OR)
• Permitted to Defend in Person or Legal Assistance • Given Legal Assistance for Free
Option of Free Access to Interpreter (AND)
• Option Provided If Needed
Allowed to Fairly Examine Witnesses (AND)
• Valid Non-attendance of Witnesses • Fairly Examine Witnesses

a chain of abstraction levels: from facts we ascribe leaf factors; from leaf factors we ascribe intermediate factors; from intermediate factors we ascribe issues; and from issues we ascribe the outcome. Hence, we omit the sections of the case summary that do not pertain to THE FACTS section.

We took inspiration from [24], whereby each ADM node targeted for ascription learning is assigned its own H-BERT model as illustrated in Figure 4. In brief, each case in the corpus is broken down into fact embeddings identified by the HTML formatting, which are then encoded according to the RoBERTa model [18], after which the encodings are fed into a BERT model to produce an overall document encoding which is passed through a feedforward neural-network (FFNN) to produce a binary classification for ascription or non-ascription of a particular node within the ADM. Note, that there will be as many BERT models with their associated feedforward network as there are ADM nodes that one wishes to ascribe. For the backwards pass during training, the RoBERTa weights are not adjusted since these are associated with fact encodings; only the BERT model and its downstream FFNN are adjusted since these are associated with the node ascription.

To capture the legal reasoning applicable to resolution of cases under Article 6 of the ECHR, we developed an ADM inspired by the ADF representation of the domain from [14]. There are four levels of abstraction within our ADM for Article 6 of the ECHR: 'Outcome' node; 'Issue' nodes; 'Intermediate' factor nodes; and 'Leaf' factor nodes. Tables 4-8 compactly represent the ADM – satisfaction of the 'Outcome' node (denoting a case results in a non-violation of the applicant's right to a fair trial) is determined as the conjunction of its five children 'Issue' nodes, that is ('FAIR' ∧ 'PUBLIC HEARING' ∧ 'REASONABLE TIME' ∧ 'PRESUMPTION OF INNOCENCE' ∧ 'HAD THE MINIMUM RIGHTS'). Satisfaction of an 'Issue' level node is determined as a conjunction of the satisfaction of its children 'Intermediate' factor nodes. Satisfaction of an 'Intermediate' factor node is determined as either a conjunction or disjunction of the satisfaction of its children 'Leaf' factor nodes as we demonstrate in Tables 4-8.

To provide classification targets for ascription, we had to cater for divergent annotations due to annotator disagreement. Fortunately, as discussed in Section 3, divergence was not a frequent problem. Annotations were interpreted as proportions to be used as probabilistic classification weights for the H-BERT models. These classification weights can also be propagated down to children nodes, where appropriate, in accordance with the acceptance logic. This weight propagation is essential to derive classification weights for 'Leaf' factor nodes, since annotations were conducted at the 'Issue' and 'Intermediate' abstraction levels.

Example 4.1. Say a particular case was annotated by five annotators, three of whom ascribed as satisfied (positive ascription) the 'Intermediate' level node **Access to Court** (under the 'Issue' node FAIR – see Table 4), one who ascribed it as unsatisfied (negative ascription), and one who did not ascribe it (non-ascription). Then its ascription weights would be the tuple [0.6, 0.2, 0.2] (respectively [positive ascription, negative ascription, non-ascription]). Recall that positive and negative ascriptions are combined for the ascription task. Hence there would be an 80% chance that the case will be input to the H-BERT model associated with the **Access to Court** node with an ascribed classification. There would also be an independent 20% chance that the case will be input to the H-BERT model with a non-ascribed classification (which means the case has

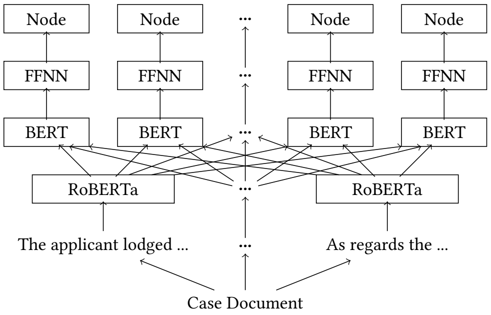


Figure 4: Feed-forward representation of the ADM H-BERT architecture for ECHR Article 6 case ascription from the input natural language document to output ascription or non-ascription of nodes contained in the ADM.

a 16% chance of being included in both the ascribed classification set and the non-ascribed classification set for **Access to Court**). Since **Access to Court** has a disjunction as its acceptance condition, it will propagate its positive ascription weight scaled by combinatorial probability (since only one child need be satisfied there are multiple combinations for satisfaction of the parent), whereas it will propagate its negative ascription weight exactly (since each child must not be satisfied for the parent not to be satisfied). The non-ascription weight is also passed exactly to its children. The combinatorial probability α is taken as $\alpha = \frac{2^{|C|-1}}{2^{|C|}-1}$, where C are the children nodes. Therefore both children (*Opportunity for Tribunal*, and *Legitimate Reasons for Limitations*) will have ascription weights [0.4, 0.2, 0.2], since $\alpha = \frac{2}{3}$ providing a 60% chance of being input with an ascribed classification, and a separate 20% chance of input with a non-ascribed classification (which means the case has a 12% chance of being included in both the ascribed classification set and the non-ascribed classification set for the two children nodes).

As is evident from Figure 2, the balance between ascription and non-ascription is heavily weighted towards non-ascription, with the notable exception of the issue 'FAIR'. We opted to oversample the less represented class to create balanced data sets whilst retaining as much information as possible, given the relatively small size of the data sets in comparison to typical NLP domains. All relevant code is available open-source[2].

5 EXPERIMENTS

In this section we assess the performance of the H-BERT models at ascribing at the three different levels of abstraction (Issue, Intermediate, Leaf). We first outline the experimental setup, providing details of the case corpus data set and describing the implementation of the relevant analysis[3], and then present the analysis and evaluation of the results.

5.1 Experimental Setup

The ADM was developed from official documentation and expert opinion pertaining to January 2015, as discussed in Section 3. However, we restricted analysis to match the data set used in [24], in order to maintain a consistent data set across research outputs. This further reduced the data set to 673 cases: 186 of the 204 non-violation and 487 of the 491 violation cases.

In conjunction with the corpus data set encodings, we have a parallel data set of case annotations produced by the law students in the research study described in Section 3. Each case in the corpus was labelled with reference to every legal node at the three levels of abstraction (Issue, Intermediate, Leaf). As outlined in Section 4, students labelled the nodes with one of three options: positive ascription; negative ascription; not ascribed. For the purpose of our analysis, we combined the positive and negative ascriptions into one joint label, so that we have a binary classification task of ascription vs non-ascription for any given legal node.

The number of instances for training and testing is smaller than the relatively vast data sets that are usually employed for NLP tasks. The corpus is used to evaluate the performance of a H-BERT approach developed specifically for small data sets [19], at ascribing at the three levels of abstraction (Issue, Intermediate, Leaf). All experiments use the same pre-trained 512 token RoBERTa model encodings, and use 256 tokens for document BERT model encoding.

The ascription task is essentially a series of binary classification tasks. A particular level of abstraction will consist of several nodes that relate to that abstraction, and each node provides its own classification task over the data set. For example, at the 'Issue' level of abstraction there are five 'Issue' nodes: General Fairness; Public Hearing; Reasonable Time; Presumption of Innocence; Had the Minimum Rights. Every case is annotated with respect to each node, providing the classification targets for the node over the data set. We repeated the classification task for each node twenty times (with set random seeds for reproducability) to provide confidence in the scope of the ascription results.

Four metrics were selected for evaluation as appropriate for the binary classification task: accuracy, macro F1 score, and MCC (Matthews correlation coefficient) score. Since the data set is unbalanced between ascription and non-ascription of nodes, as depicted in Figure 2, the F1 and the MCC scores are of clear relevance. We explicitly include the F1 score for ascription (denoted as Ascribe F1 in Table 9), since reliable ascription of legal nodes is arguably more important than reliable non-ascription. The importance of the Ascribe F1 metric will be further discussed in Section 6. We normalise all metrics to the ranges [-100, 100] or [0, 100] in order to provide a consistent presentation.

A total of 1140 experiments were conducted; 20 experiments for each legal node at each level of abstraction. Specifically, the 'Issue' abstraction level contained 5 nodes, the 'Intermediate' level contained 18 nodes, and the 'Leaf' level contained 34 nodes, giving 57 nodes in total. Each experiment was randomly split, by defined seeds for reproducability, into 80% training and 20% test data, and underwent 30 epochs of training. Two Nvidia Tesla P100 GPUs were used for fine-tuning.

[2] Code available at https://github.com/jamumford/ECHR_Article6_ADM_Ascribe
[3] Undertaken on Barkla – High Performance Computing facilities, at the University of Liverpool, UK.

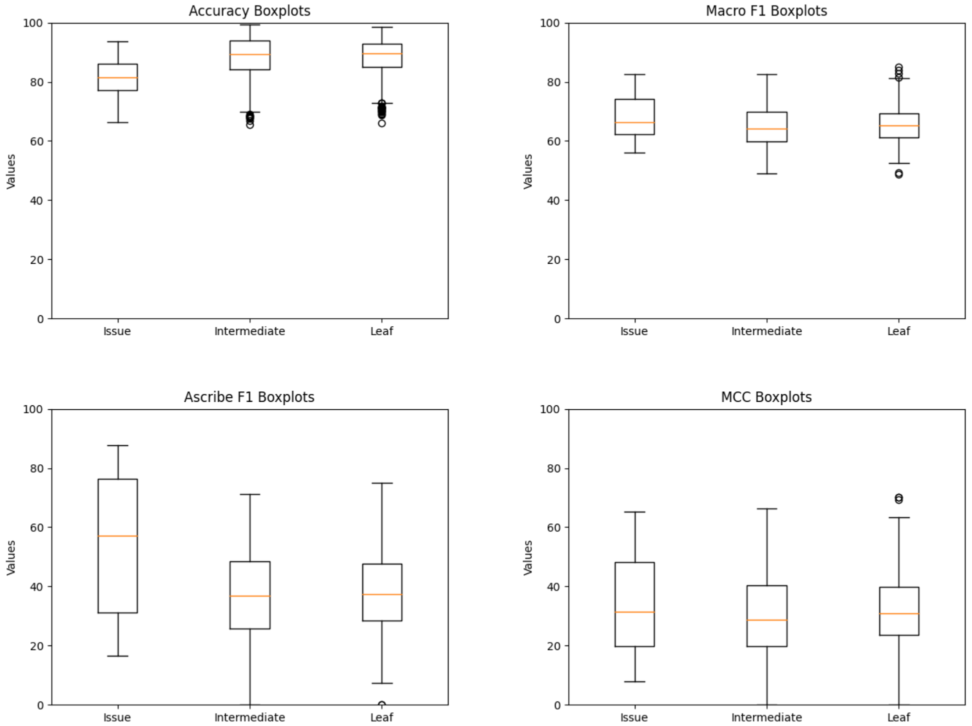


Figure 5: Box plot illustrations of the distributions achieved by H-BERT models on test data sets for ascription at the three abstraction levels ('Issue', 'Intermediate', 'Leaf') across the metrics: accuracy, ascribe F1 score, macro F1 score, and MCC score.

5.2 Results and Evaluation

The experimental results are summarised in Table 9 and provide a comparison of the performance of the H-BERT models over the three levels of abstraction. With the exception of the accuracy metric, the H-BERT model trained for ascription at the 'Issue' level produces the best results. The Mann-Whitney p values indicate that most comparison inferences are statistically significant. For example, a p value less than 0.0001 would indicate that the hypothesis that the distributions for two particular abstraction levels belong to the same population can be rejected at the 99.99% confidence level.

The results for Sets 1 and 2 indicate that we can be confident in the superior returns in F1 scores produced at the 'Issue' level of abstraction, as well as the inferior returns produced in terms of accuracy. We can be less confident about the differences in MCC scores, although the superior performance over the 'Intermediate' level of abstraction (Set 1) is associated with a p value that provides a reasonable degree of confidence, the p value associated with comparison against the 'Leaf' level of abstraction (Set 2) is too high for us to be confident that the distributions are distinct.

The results for Set 3 provide p values that offer little confidence that the distributions can be viewed as distinct in terms of performance under the accuracy and ascribe F1 metrics. However, the p value less than 0.05 for macro F1 suggests we can reject the hypothesis of shared populations at the 95% confidence level, indicating improved performance at the 'Leaf' abstraction level over the 'Intermediate' level. Furthermore the p value for the MCC metric is sufficiently low such that we can confidently consider the superior performance at the 'Leaf' level to be statistically significant in terms of capturing the overall distribution of ascription classifications.

The comparison offered in Tables 10-11 support the effectiveness of the ADM for representing the relevant domain knowledge. The 'random' and 'majority class' classifiers are the two classic 'sanity' tests for evaluating the scope of a problem since they are considered 'dumb' approaches. Results for these two classifiers were clearly outperformed by the relevant H-BERT model at each abstraction level. The only exception to this pattern was with respect to the accuracy returns at the 'Intermediate' and 'Leaf' abstraction levels, where the 'majority class' classifier achieved slightly higher mean returns. However, this serves to further highlight the poor relevance of the accuracy metric at indicating the performance of a classifier operating on an unbalanced data set. The near zero MCC score returns for the 'random' and 'majority class' classifiers support the use of the MCC score as the strongest indicator of effective performance.

Across all metrics and all abstraction levels there is clearly a great degree of variance of results as indicated by the range for each return illustrated in Figure 5. This variance is not due to anomalies, nor is it proportional in the quantity of nodes at the abstraction level since the 'Issue' level has similar ranges to the other two levels and consists of far fewer nodes. Rather, **the variance would appear to result from the uneven distribution of ascription vs non-ascription over the nodes.** Those nodes having a greater balance of ascription and non-ascription in the annotated data set tend to offer higher returns over the metrics, whereas those nodes with sparse ascription provide low returns. It is notable that some nodes, at each abstraction level, are associated with exceptional performance across the metrics, whereas others were associated with very poor performance.

Table 9: Comparison of ascription performance of H-BERT models trained on the three levels of abstraction (Issue, Intermediate, Leaf) of annotated data sets of cases pertaining to ECHR Article 6. We provide the means and the upper and lower ranges (2 d.p.). Mann-Whitney p values (3 d.p.) are indicated by Set n, where: Set 1 is the comparison between returns for the Issue and Intermediate abstraction levels; Set 2 is between the Issue and Leaf abstraction levels; and Set 3 is between the Intermediate and Leaf abstraction levels.

	Accuracy	Ascribe F1	Macro F1	MCC
Issue	$81.03^{+12.72}_{-14.62}$	$\mathbf{54.97^{+32.75}_{-38.30}}$	$\mathbf{67.99^{+14.68}_{-12.04}}$	$\mathbf{34.15^{+31.19}_{-26.16}}$
Inter	$\mathbf{88.21^{+11.01}_{-22.80}}$	$36.56^{+34.55}_{-36.56}$	$64.90^{+17.74}_{-15.87}$	$30.01^{+36.26}_{-30.01}$
Leaf	$88.19^{+10.23}_{-22.04}$	$38.18^{+36.82}_{-38.18}$	$65.74^{+19.33}_{-16.90}$	$32.09^{+38.06}_{-32.09}$
Set 1	0.0000	0.0000	0.0004	0.0377
Set 2	0.0000	0.0000	0.0070	0.3182
Set 3	0.2287	0.1149	0.0340	0.0085

Table 10: Comparison of ascription performance of a random classifier on the three levels of abstraction of annotated data sets of cases pertaining to ECHR Article 6.

	Accuracy	Ascribe F1	Macro F1	MCC
Issue	$50.01^{+0.03}_{-0.06}$	$32.77^{+26.43}_{-19.26}$	$44.54^{+4.03}_{-5.41}$	$0.18^{+0.40}_{-0.14}$
Inter	$50.01^{+0.12}_{-0.13}$	$16.19^{+18.26}_{-12.47}$	$40.14^{+6.86}_{-5.18}$	$0.07^{+0.11}_{-0.07}$
Leaf	$50.00^{+0.16}_{-0.16}$	$14.77^{+17.53}_{-11.06}$	$39.58^{+6.78}_{-4.64}$	$0.06^{+0.14}_{-0.04}$

Table 11: Comparison of ascription performance of a majority-class classifier on the three levels of abstraction of annotated data sets of cases pertaining to ECHR Article 6.

	Accuracy	Ascribe F1	Macro F1	MCC
Issue	$78.94^{+12.87}_{-14.27}$	$16.79^{+67.17}_{-16.79}$	$43.92^{+3.93}_{-4.69}$	$0.00^{+0.00}_{-0.00}$
Inter	$89.04^{+8.32}_{-16.36}$	$0.00^{+0.00}_{-0.00}$	$47.03^{+2.30}_{-4.97}$	$0.00^{+0.00}_{-0.00}$
Leaf	$90.04^{+6.44}_{-13.56}$	$0.00^{+0.00}_{-0.00}$	$47.32^{+1.78}_{-4.01}$	$0.00^{+0.00}_{-0.00}$

Each H-BERT experiment required roughly 5.5 minutes of training for a given node over the full data set. Therefore, training time for 20 experiments on each node took: 9 hours 32 minutes at the 'Issue' abstraction level; 37 hours 14 minutes at the 'Intermediate' level; and 59 hours 28 minutes at the 'Leaf' level. It is important to note that once trained, employing any learned H-BERT model on a new case requires negligible time to produce its ascription classification and attention weights over the facts of the case. Changes to the law are also gradual, with a slow rate of generation of new resolved cases (i.e., new data) that renders any fine-tuning exercise on new data very manageable.

6 DISCUSSION

It is uncontroversial that any practical AI system in the legal domain must be able to explain and justify its outputs. Other domains that are less prone to risk, may need not be held to the same standard and may perhaps be judged on performance metrics alone. We would argue that any data-driven method that leaps straight from natural language description to case outcome classification/prediction without adherence to the reasoning used in the legal domain, will be severely limited in its ability to be adequately audited to meet appropriate standards (and in some cases, regulations, such as the European GDPR) of explainability and justifiability. By instead using data-driven techniques to ascribe the input to an approved domain model that captures the relevant legal reasoning, one may strike a balance between performance metrics and such auditing. Furthermore, recent strides in attempting to apply the impressive gains in NLP performance from LLMs (large language models) [16] to the legal domain have not in general met with success similar to that evident in less complicated domains. Results described in papers such as [13] and [15] point to the limitations of transformer-based approaches, offering performances difficult to distinguish from earlier ML approaches such as [4]. These observations are well supported in [21], where three versions of outcome prediction are identified: *outcome identification*, *outcome-based judgement categorisation* and *outcome forecasting*. The aforementioned papers applying LLMs are instances of outcome identification, which is argued in [21] to be an unsound choice for application to ECtHR judgement cases, which are always provided pre-labelled with their outcome. The approach presented in this paper falls into their definition of outcome-based judgement categorisation, with ascription to the ADM used to explain and justify legal case outcomes.

The relevance of factors for explanation and justification when processing legal cases has been advocated in [9] and [25]. In [23] the argument was made that when reasoning with legal cases, ascription from the fact level to the factor level is the role for which ML is most suitable. We can see from the results of the previous section, and Table 9 in particular, that ascription at the 'Issue' abstraction level offered the highest return for all performance metrics with the exception of accuracy. But if we consider the shape of the data set in terms of balance between the distribution of classes (ascribed vs non-ascribed), then we would expect the 'Issue' abstraction level to produce the best results due to its greater balance between the classes (see Figure 2) since ascription is shared between only five nodes. When we consider the dilution of ascription that occurs by the time we are considering the 34 'Leaf' factor nodes, the corresponding returns on the metrics are still impressively high. This is especially true for the MCC scores, which are statistically inseparable from the 'Issue' MCC scores with a significant degree of confidence, and indicate the models are effective at capturing the distribution of the binary ascription classes.

Furthermore, classification learning at the 'Issue' and 'Intermediate' levels of abstraction benefits from data sets derived from the direct labelling produced by the annotators in the research study described in Section 3. The data sets used for classification learning at the 'Leaf' abstraction level, were instead indirectly derived by propagating weights down from the 'Intermediate' level with a rigid combinatorial probability assumption that has no guarantees of legitimacy. Therefore, that the ascription performance across the metrics, but especially for MCC, is relatively high and even strong for some nodes, is an encouraging output. The results support further investigation into the extent to which the H-BERT models actually ascribe for the right reasons, with particular focus on the nodes at the extreme ends of the performance ranges. High levels of performance by an ML system are no guarantee that it is applying the correct rationale, as demonstrated in [28]. For each node, our implementation was designed to store the most successful H-BERT model across all the runs, and we intend to examine the attention weights of these models as a potentially fruitful research direction.

7 CONCLUSION

In this paper we have outlined our approach to ascription of factors and issues to legal cases by combining a symbolic domain model, called an ADM (Angelic Domain Model), with a state-of-the-art H-BERT (Hierarchical Bidirectional Encoder Representations from Transformers) NLP technique. We developed an ADM to capture the reasoning and knowledge relevant to Article 6 (the right to a fair trial) of the ECHR (the European Convention of Human Rights). This ADM was used as part of a research study that involved the annotation of a corpus of cases pertaining to Article 6, enabling us to obtain a clear perspective on the distribution of the factors relevant to the complaint of a potential violation of Article 6. Most importantly, the research study provided an annotated data set for training H-BERT models to ascribe in accordance with the ADM. A series of experiments were conducted to evaluate the effectiveness of the H-BERT models at the ascription task at the different abstraction levels of the ADM. The results of these experiments are very encouraging and support the suitability for machine learning to be directed at the task of factor ascription. Our future work will investigate the attention weights associated with the H-BERT models, to evaluate if the models can highlight relevant passages to support their ascriptions in a justifiable manner. We will also evaluate the H-BERT models against benchmarks of alternative ML approaches that can highlight relevant text for justified ascription, having laid the foundations in this paper. Furthermore, we plan to evaluate and refine the annotated data set in accordance with domain expert judgement, in order to establish a reliable gold standard for training and testing ML classifiers on the ascription task.

ACKNOWLEDGMENTS

This work was supported by Towards Turing 2.0 under the EPSRC Grant D-ACAD-052 & The Alan Turing Institute.

REFERENCES

[1] Latifa Al-Abdulkarim, Katie Atkinson, and Trevor Bench-Capon. 2016. Accommodating change. *AI and Law* 24, 4 (2016), 409–427.

[2] Latifa Al-Abdulkarim, Katie Atkinson, and Trevor Bench-Capon. 2016. A methodology for designing systems to reason with legal cases using ADFs. *AI and Law* 24, 1 (2016), 1–49.

[3] Latifa Al-Abdulkarim, Katie Atkinson, Trevor Bench-Capon, Stuart Whittle, Rob Williams, and Catriona Wolfenden. 2019. Noise induced hearing loss: Building an application using the ANGELIC methodology. *Argument & Computation* 10, 1 (2019), 5–22.

[4] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2 (2016), e93.

[5] Vincent Aleven. 1997. *Teaching case-based argumentation through a model and examples.* Ph.D. thesis. University of Pittsburgh.

[6] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. *Artificial Intelligence* 289 (2020), 103387.

[7] Trevor Bench-Capon. 2021. Using issues to explain legal decisions. In *eXplainable and Responsible AI and Law 2021*, Vol. 3168 CEUR Workshop Proceedings. 1–22.

[8] Trevor Bench-Capon and Thomas F. Gordon. 2022. Implementing a Theory of a Legal Domain. In *Proceedings of JURIX 2023.* 13–22.

[9] L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *AI and Law* 29, 2 (2021), 213–238.

[10] Gerhard Brewka, Hannes Strass, Stefan Ellmauthaler, Johannes Peter Wallner, and Stefan Woltran. 2013. Abstract dialectical frameworks revisited. In *Twenty-Third International Joint Conference on Artificial Intelligence.* 803–809.

[11] Gerhard Brewka and Stefan Woltran. 2010. Abstract dialectical frameworks. In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning.* 102–111.

[12] Stephanie Brüninghaus and Kevin D Ashley. 2003. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law.* ACM, 233–242.

[13] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

[14] Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2023. Explainable AI Tools for Legal Reasoning about Cases: A Study on The European Court of Human Rights. *Artificial Intelligence* (2023), 103861.

[15] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683* (2022).

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77, 2 (1995), 321–357.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. A sentence-level hierarchical bert model for document classification with limited labelled data. In *Proceedings of DS 2021.* Springer, 231–241.

[20] Masha Medvedeva, Michel Vols, and Martijn Wieling. 2019. Using machine learning to predict decisions of the European Court of Human Rights. *AI and Law* (2019), 1–30.

[21] Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *AI and Law* 31, 1 (2023), 195–212.

[22] Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2021. Explaining Factor Ascription. In *Proceedings of JURIX 2021.* 191–196.

[23] Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2021. Machine learning and legal argument. In *Computational Models of Natural Argument 2022*, Vol. 2937 CEUR Workshop Proceedings. 47–56.

[24] Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2022. Reasoning with Legal Cases: A Hybrid ADF-ML Approach. In *Proceedings of Jurix 2022.* 93–102.

[25] Henry Prakken and Rosa Ratsma. 2021. A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument & Computation* (2021), 1–36.

[26] Henry Prakken and Giovanni Sartor. 1998. Modelling Reasoning with Precedents in a Formal Dialogue Game. *AI and Law* 6, 2-3 (1998), 231–287.

[27] Edwina L Rissland and Kevin D Ashley. 1987. A case-based system for Trade Secrets law. In *Proceedings of the 1st International Conference on Artificial Intelligence and Law.* 60–66.

[28] Cor Steging, Silja Renooij, and Bart Verheij. 2021. Discovering the rationale of decisions: towards a method for aligning learning and reasoning. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law.* 235–239.

[29] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine* 37, 5 (2005), 360–363.

[30] Serena Villata, Michal Araszkiewicz, Kevin D Ashley, Trevor Bench-Capon, L Karl Branting, Jack G Conrad, and Adam Wyner. 2022. Thirty years of *Artificial Intelligence and Law*: the third decade. *AI and Law* 30, 4 (2022), 561–591.