AI & Law: Formative Developments, State-of-the-Art Approaches, Challenges & Opportunities

Jack G. Conrad
jack.g.conrad@thomsonreuters.com
TR Labs at Thomson Reuters
Minneapolis, MN, USA

Shirsha Ray Chaudhuri
Shirsha.RayChaudhuri@thomsonreuters.com
TR Labs at Thomson Reuters
Bangalore, Karnataka, India

Shounak Paul
shounakpaul95@kgpian.iitkgp.ac.in
IIT Kharagpur
Kharagpur, West Bengal, India

Saptarshi Ghosh
saptarshi@cse.iitkgp.ac.in
IIT Kharagpur
Kharagpur, West Bengal, India

ABSTRACT

Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) are transforming the way legal professionals and law firms approach their work. The significant potential for the application of AI to Law, for instance, by creating computational solutions for legal tasks, has intrigued researchers for decades. This appeal has only been amplified with the advent of Deep Learning (DL). In particular, research in AI & Law can be extremely beneficial in countries like India with an overburdened legal system.

In this tutorial, we will give an overview of the various aspects of applying AI to legal textual data. We will start with a history of AI & Law, and then discuss the current state of AI & Law research including the techniques that have produced the biggest impact. We will also take a deep dive into the software processes required to implement and sustain such AI solutions.

KEYWORDS

Legal Analytics, Natural Language Processing, Text Analytics, Machine Learning

ACM Reference Format:
Jack G. Conrad, Shirsha Ray Chaudhuri, Shounak Paul, and Saptarshi Ghosh. 2023. AI & Law: Formative Developments, State-of-the-Art Approaches, Challenges & Opportunities. In *6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023), January 4–7, 2023, Mumbai, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3570991.3571050

1 DESCRIPTION & OUTLINE

This tutorial will attempt to give an overview of different aspects of applying AI techniques, especially Natural Language Processing and Machine Learning, to legal textual data. It is worth noting that working with legal text is far more challenging than in many other subdomains of NLP, mainly due to factors like lengthy documents, complex language and lack of large-scale datasets.

The tutorial will have three parts (along with Q&A), as summarized in Table 1. The three parts are detailed below.

- **Part 1:** *A brief history of AI & Law research, important milestones*

  The tutorial will start with a brief history of the application of AI in the legal domain, and its progress over the decades, from rudimentary approaches to sophisticated Deep Learning technologies [1, 7–9, 16, 17]. We shall also discuss the inherent challenges of AI-based approaches for solving practical tasks on legal text.

- **Part 2:** *State-of-the-art in AI & Law research, datasets, benchmarks and tools*

  We shall then discuss some of the latest AI & Law research challenges that are being pursued in India (and globally) today, and the state-of-the-art techniques being used to address these challenges. In particular, we shall discuss problems such as semantic segmentation of legal documents [4], legal statute identification from fact descriptions [14], court judgment prediction [12], and the summarization of legal documents [2]. We shall also introduce publicly available resources that are useful for research in AI & Law, such as datasets [4, 12, 14], benchmarks [6] and pre-trained language models (PLMs) [5, 10, 15, 19]. Table 2 summarizes some of the datasets/resources that we plan to discuss.

- **Part 3:** *Challenges and best practices in developing Legal information systems at scale (how to put the AI models into practice and sustain them)*

  Imagine that the extensive and rigorous research that we describe in Part 2 is now completed. We will consider in this part, technical challenges, and the software development life-cycle that is required to implement a software solution that allows the expected consumption of the model thus developed. We will talk about the tech stack required to enable state-of-the-art models in public clouds like AWS and the overlap with the ML life-cycle in a sustainable manner.

Note that research in AI & Law is not only limited to processing of English legal text. A large number of prior studies have worked on legal text in other European languages [11], Chinese legal text [18],

CODS-COMAD 2023, January 4–7, 2023, Mumbai, India

ACM ISBN 978-1-4503-9797-1/23/01…$15.00
https://doi.org/10.1145/3570991.3571050

Part	Topic	Presenter	Duration
1	A brief history of AI & Law research, important milestones	Jack Conrad	30 mins
2	State-of-the-art in AI & Law research, datasets, benchmarks and tools	Saptarshi Ghosh, Shounak Paul	25 mins
3	Challenges and best practices in developing Legal information systems at scale (how to put the AI models into practice and sustain them)	Shirsha Ray Chaudhuri	25 mins
4	Q&A	All Presenters	10 mins

Table 1: Brief outline of the entire tutorial session (duration 90 minutes)

Datasets	
Dataset	**Description**
Semantic Segmentation [4]	**Corpus of 150 Indian Supreme Court Case Documents annotated for the Semantic Segmentation Task:** The task involves splitting a case document into functional parts (Facts, Arguments, Ruling by Present Court, etc.) by labeling individual sentences under a sentence-tagging framework. 7 semantic segments have been considered and annotated by experts in this dataset.
ILSI [14]	**Corpus of 65k Indian criminal court case documents for the Legal Statute Identification Task:** The task requires one to find out the relevant statutes (written law articles) that can be relevant given the facts of a case, using a multi-label text classification framework. The label set consists of the 100 most frequent statutes from the Indian Penal Code.
ILDC [12]	**Multiple corpora of Indian Supreme Court case documents for the Court Judgment Prediction and Explanation Task:** The task involves predicting whether the claims made in the case are accepted/rejected under a Binary Text Classification Framework. *ILDC-single* (6k docs) contain a single claim per case, *ILDC-multi* (35k docs) contain multiple claims per case and there is also a small set of docs (50) having human annotations for explainability.
LexGLUE Benchmark [6]	**A collection of multiple datasets (mostly EU, UK or US-based) for different legal tasks:** Comprises of tasks such as binary and multi-label statute identification (ECtHR-A and ECtHR-B), predicting relevant legal issues (SCOTUS) or concepts (EUR-LEX), predicting unfair terms-of-service (UNFAIR-ToS) and MCQ answering regarding case holdings (CaseHOLD).
Pre-trained Language Models	
Model	**Description**
LegalBERT [5]	**Pre-trained Language Model over EU, UK and US legal text:** Pre-trained *BERT-base-uncased* from scratch (using custom tokenizer) over a corpus of 350K documents (12GB) comprising of the legislation of EU and UK, cases from the ECJ and ECHR, and case documents from the US.
CaseLawBERT [19]	**Pre-trained Language Model over US case documents:** Pre-trained *BERT-base-uncased* both from scratch (using custom tokenizer) and continual training over US State and Federal court case documents (3.4M docs, 37 GB) from the Harvard Case Law Project.
PoLBERT [10]	**Pre-trained Language Model over many types of legal documents:** Pre-trained *BERT-base-large* using two different seeds with the RoBERTa pre-training objective over a huge corpus of legal documents (10M docs, 256 GB), such as legal analyses, court opinions, government publications, contracts, statutes, regulations, and more, mostly from the US and EU.
InLegalBERT & InCaseLawBERT [15]	**Pre-trained Language Model over Indian court case documents:** Pre-trained *BERT-base-uncased* over a corpus of Indian statutes and case documents (5.4M, 27GB) by continuing training the LegalBERT and CaseLawBERT models above.

Table 2: Brief Description of some of the resources to be discussed in Part 2 of the tutorial

and so on. But in this tutorial, we will only focus on application of AI on English legal text.

2 GOALS / OBJECTIVES OF THE TUTORIAL

Through this tutorial, the participants will become familiar with various challenges and opportunities in the field of AI & Law, which is an emerging research area for AI and NLP researchers. Additionally, there is a critical need for AI & Law solutions in many countries where the legal system is highly overburdened and access to justice is costly and burdensome for the common citizen.

With countries across the world making efforts in digitizing legal records and funding research in AI & Law, the future is bright and the potential scope is huge. Furthermore, AI & Law has made

significant inroads in industry as well, being increasingly adopted by law firms and corporations, including startups, to introduce cutting-edge solutions. Hence the research community will benefit from this tutorial.

3 TARGET AUDIENCE

Anyone interested in AI & Law will benefit from this tutorial. In particular, we believe that this tutorial will benefit technologists who wish to apply AI in the legal domain, prospective AI/NLP researchers, as well as legal practitioners interested in contemporary technological solutions to legal problems. We would also love to have tech enthusiasts interested in the application of AI to legal content attend.

A knowledge of NLP and standard Machine Learning techniques would be helpful for attendees to grasp the NLP tools covered during the tutorial.

4 PRESENTERS OF THE TUTORIAL

- **Jack G. Conrad**, *Director of Applied Research and Lead Research Scientist, TR Labs at Thomson Reuters, Minneapolis, MN USA*

  Jack Conrad is Director of Applied Research at Thomson Reuters TR Labs where he focuses on a broad range of technical application areas involving AI, machine learning and textual data processing. He also fosters cross-team collaboration and communication in the process of implementing and deploying technology to meet business needs. For over two decades, he has delivered critical artifacts and infrastructure for research and business directed projects across a diverse spectrum of domains that have included legal, tax and news. Jack has published more than 50 peer reviewed research papers and has eight patents. He is passionate about the power of AI transformation in enterprise environments.
  Jack is past president of the International Association for Artificial Intelligence and Law (IAAIL.org) and has served on the IAAIL Executive Committee for 8 years. Jack's areas of expertise include research in the fields of information retrieval (search), question answering, NLP, machine learning, data mining, and system evaluation.

- **Shirsha Ray Chaudhuri**, *Director of Engineering, TR Labs at Thomson Reuters, Bangalore, Karnataka, India*

  Shirsha Ray Chaudhuri is the Director of Engineering at Thomson Reuters Labs, Bangalore. TR Labs is TR's applied research division, focused on delivering solutions with AI and emerging tech to TR's platforms and products and customer Proof of Concepts (PoCs). TR's editorial workflows power its best-in-class product like Westlaw. The TR Labs team in Bangalore contributes to leveraging AI in these editorial workflows. Her earlier work includes strategic architecture and prototyping for Daimler's EvoBus team to help route planning software solutions for electric buses, predictive maintenance solutions for Daimler's DTNA service centres, and use of AI for rapid field ops in Daimler's Japan-based FUSO trucks. Besides providing point-in-time solutions for a single use case, she implemented generic modifications of such AI services which could be replicated across geographies and operators.

- **Shounak Paul**, *Senior Research Fellow, Deptt. of Computer Science &, Engineering, IIT Kharagpur, West Bengal, India*

  Shounak Paul is a Senior Research Fellow at the Department of CSE, IIT Kharagpur. His research interests mainly include legal data analytics and applications of NLP in the legal domain. His works on AI & Law for Indian applications have been published in premier conferences and journals such as: semantic segmentation [3, 4] (JURIX 2019, best paper award; AI & Law Journal 2021), charge identification [13] (COLING 2020) and legal statute identification using citation networks [14] (AAAI 2022).

- **Saptarshi Ghosh**, *Assistant Professor, Deptt. of Computer Science &, Engineering, IIT Kharagpur, West Bengal, India*

  Saptarshi Ghosh is an Assistant Professor at the Department of CSE, IIT Kharagpur. His research interests include Legal analytics, Social media analytics, and Algorithmic bias and fairness (on which he presently leads a Max Planck Partner Group at IIT Kharagpur). His works on AI & Law have been published at premier conferences including SIGIR, AAAI, CIKM, ECIR, COLING, and have been awarded at top AI & Law conferences, including the *Best Paper Award* at the International Conference on Legal Knowledge and Information Systems (JURIX) 2019, and the *Best Student Paper Award* at the International Conference on Artificial Intelligence and Law (ICAIL) 2021. He is presently the Section Editor on Legal Information Retrieval for the journal *Artificial Intelligence and Law*, the premier journal in AI & Law. He has been awarded with several prestigious awards, including the Institution of Engineers (India) Young Engineer Award 2017-18 in Computer Engineering discipline.

REFERENCES

[1] Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319. https://link.springer.com/article/10.1007/s10506-012-9131-x

[2] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *European Conference on Information Retrieval*. Springer, 413–428. https://link.springer.com/chapter/10.1007/978-3-030-15712-8_27

[3] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, Vol. 322. IOS Press, 3. https://arxiv.org/abs/1911.05405

[4] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2021. DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law* (2021), 1–38. https://link.springer.com/article/10.1007/s10506-021-09304-5

[5] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[6] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers). Association for Computational Linguistics, Dublin, Ireland, 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

[7] Jack G. Conrad and John Zeleznikow. 2013. The Significance of Evaluation in AI and Law: A Case Study Re-Examining ICAIL Proceedings. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law* (Rome, Italy) *(ICAIL '13)*. Association for Computing Machinery, New York, NY, USA, 186–191. https://doi.org/10.1145/2514601.2514624

[8] Jack G. Conrad and John Zeleznikow. 2015. The Role of Evaluation in AI and Law: An Examination of Its Different Forms in the AI and Law Journal. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (San Diego, California) *(ICAIL '15)*. Association for Computing Machinery, New York, NY, USA, 181–186. https://doi.org/10.1145/2746090.2746116

[9] Guido Governatori, Trevor Bench-Capon, Bart Verheij, Michał Araszkiewicz, Enrico Francesconi, and Matthias Grabmair. 2022. Thirty years of Artificial Intelligence and Law: the first decade. *Artificial Intelligence and Law* (2022), 1–39. https://link.springer.com/article/10.1007/s10506-022-09329-4

[10] Peter Henderson, Mark S Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. *arXiv preprint arXiv:2207.00220* (2022). https://arxiv.org/abs/2207.00220

[11] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A Dataset of German Legal Documents for Named Entity Recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. 4478–4485.

[12] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4046–4062. https://doi.org/10.18653/v1/2021.acl-long.313

[13] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2020. Automatic Charge Identification from Facts: A Few Sentence-Level Charge Annotations is All You Need. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1011–1022. https://doi.org/10.18653/v1/2020.coling-main.88

[14] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2022. LeSICiN: A Heterogeneous Graph-Based Approach for Automatic Legal Statute Identification from Indian Legal Documents. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (Jun. 2022), 11139–11146. https://doi.org/10.1609/aaai.v36i10.21363

[15] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pretraining Transformers on Indian Legal Text. *arXiv preprint arXiv:2209.06049* (2022). https://arxiv.org/abs/2209.06049

[16] Giovanni Sartor, Michał Araszkiewicz, Katie Atkinson, Floris Bex, Tom van Engers, Enrico Francesconi, Henry Prakken, Giovanni Sileno, Frank Schilder, Adam Wyner, et al. 2022. Thirty years of Artificial Intelligence and Law: the second decade. *Artificial Intelligence and Law* (2022), 1–37. https://webspace.science.uu.nl/~prakk101/pubs/SecondDecadeComplete.pdf

[17] Serena Villata, Michal Araszkiewicz, Kevin Ashley, Trevor Bench-Capon, L Karl Branting, Jack G Conrad, and Adam Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law* (2022), 1–31. https://link.springer.com/article/10.1007/s10506-022-09327-6

[18] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open* 2 (2021), 79–84.

[19] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 159–168. https://arxiv.org/abs/2104.08671