Legal IR and NLP: the History, Challenges, and State-of-the-Art

Debasis Ganguly[1], Jack G. Conrad[2], Kripabandhu Ghosh[3], Saptarshi Ghosh[4], Pawan Goyal[4], Paheli Bhattacharya[4], Shubham Kumar Nigam[5], and Shounak Paul[4]

[1] University of Glasgow, United Kingdom
[2] TR Labs, Thomson Reuters, Minneapolis MN, USA
[3] Indian Institute of Science Education And Research Kolkata, India
[4] Indian Institue of Technology Kharagpur, India
[5] Indian Institute of Technology Kanpur, India

Abstract. Artificial Intelligence (AI), Machine Learning (ML), Information Retrieval (IR) 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). It is worth noting that working with legal text is far more challenging than in many other subdomains of IR/NLP, mainly due to factors like lengthy documents, complex language and lack of large-scale datasets.

In this tutorial, we introduce the audience to the nature of legal systems and texts, and the challenges associated with processing legal documents. We touch upon the history of AI and Law research, and how it has evolved over the years from rudimentary approaches to DL techniques. We give a brief introduction into the state-of-the-art research in general domain IR and NLP. We then discuss in more detail about specific IR/NLP tasks in the legal domain and their solutions, available tools and datasets, as well as the industry perspective. The discussions are followed by a hands-on coding/demo session.

Keywords: AI & Law · Legal Data Analytics · Natural Language Processing · Legal Information Retrieval.

1 Goals / Objectives of the Tutorial

AI & Law is one of the most promising areas in IR/NLP research, having become very popular in the research community over the past decade (as is evident from the increasing volume of research works in this field). Additionally, the critical need for AI and Law solutions in many countries with highly overburdened legal systems and where access to justice is costly and cumbersome for the common citizen, is motivating researchers to contribute to this field having such great degree of practical utility.

Due to the necessities listed above, countries across the world are making efforts in digitizing legal records and funding research in AI and Law. Thus, the future is promising in terms of availability of resources and funds, and the potential scope of automating practical problems in the legal domain is huge. Furthermore, applications of AI & Law are increasingly being adopted in industrial settings as well; law firms and

Table 1. Plan of the tutorial

Topic	Presenter(s)
Background on Legal Text Challenges in processing legal text	Saptarshi, Kripa, Debasis
A brief history of AI & Law research, important milestones	Jack
Background on recent approaches in NLP & IR	Pawan
Survey on the state-of-the-art in Legal IR/NLP problems; available tools and datasets	Saptarshi, Paheli, Debasis
The industry perspective	Jack
Future directions in IR/NLP/ML research in legal domain	Jack, Kripa
Hands-on Coding / demo session	Debasis, Pawan, Paheli, Shounak, Shubham

startups are investing significant effort and resources to introduce cutting-edge solutions to many of such problems. Thus, research in IR/NLP based AI & Law is likely to be a lucrative opportunity in both academia and industry alike.

Through this tutorial, the participants are likely to gain some understanding of the challenges and opportunities in this field, which are quite unique as compared to other sub-domains of IR/NLP. Although there have been some tutorials in this field in recent years, such as in ICAIL 2017 [6] and IJCAI-ECAI 2018 [7], the topics to be covered in this tutorial are significantly different. Moreover, there has not been any tutorial on this important domain in prominent IR/NLP conferences of late.

Target audience: Anyone interested in AI & Law would benefit from this tutorial. Tech enthusiasts who are interested in the practical application of IR/NLP/ML to sub-domains like Law, can also benefit from this tutorial.

Prerequisites: A knowledge of basic IR, NLP and ML techniques would be helpful for attendees to grasp the tasks and tools discussed in this tutorial; hence we view the tutorial to be at the 'intermediate' level.

2   Tutorial Outline

While there have been lot of work on legal text in other European languages [23] and in Chinese [36], this tutorial primarily focuses on application of AI on English legal text. The brief outline of the tutorial format is provided in Table 1.

Topic 1 – Background on legal text: The tutorial starts with a brief introduction to the basics of legal systems used across the world (Civil and Common Law Systems), and different types of legal documents (e.g., court case documents, statutes, patents, contracts, etc.). There is also a brief discussion on the *legal citation networks* between court case documents, statutes, etc.

---

[6] https://www.andrew.cmu.edu/user/mgrabmai/ainlawtutorial2017/
[7] https://www.ijcai-18.org/wp-content/uploads/2018/05/
    T04-AI-and-the-Law-IJCAI-ECAI-18.pdf

Topic 2 – Challenges in processing legal text: Legal documents are quite distinct from those in the open domain, such as being lengthy or being written in formal, complex language. Tasks considered relatively simple in the open domain, such as detecting sentence boundaries or Named Entity Recognition (NER), are challenging in the legal domain [32]. Moreover, in many countries, legal documents are unstructured (do not have demarcations such as headings, etc.), and do not follow any common pattern.

Although it is now possible to gain access to unlabelled legal corpora relatively easily, large-scale, task-specific annotated datasets are still difficult to find for many countries/languages. Existing datasets are not well organized, and manual annotation requires the assistance of legal experts, which proves to be expensive both in terms of time and effort. These challenges are discussed in this part.

Topic 3 – A brief history of AI & Law research, and important milestones: This part examines the roots of AI and Law, noting that this field has been around for much longer than the recent focus would suggest. We explore the field in terms of its development and expansion starting in the 1980s and study how seminal research was conducted and reported on in conference proceedings such as ICAIL and publications such as the AI and Law journal [19,31,35,2,13,14,12]. After this look at its foundations, we look at the more recent history and examine use-cases and AI-based applications that have been created to address them.

Topic 4 – Brief background of recent approaches in NLP and IR: This section briefly discusses the recent advances in NLP and IR that are particularly useful in the legal domain, such as architectures like transformers [34], and resources such as pre-trained embeddings and language models, such as BERT [15], RoBERTa [25] and XLNet [37]. We also touch upon advances in practical use cases of NLP such as low resource settings [30,16,22], and so on.

Topic 5 – A survey on the state-of-the-art in some specific problems in Legal IR/NLP, available tools and datasets: Here we discuss some common tasks and approaches in the legal domain. These include tasks like summarization [5,33], legal judgment prediction [26] and semantic segmentation [6,7], which are typically solved through approaches like retrieval (recall-oriented search), text classification (both sentence and document-level), and so on. In the legal domain, often alternative sources of knowledge (such as citation networks) can be combined with classical text processing for tasks like document similarity [4] and legal statute identification [28]. Also, recently, many pre-trained models for the legal domain have been developed [10,39,20,29], which can be applied with good results on many end tasks. Table 2 states some of the datasets and tools to be covered. We also briefly cover popular shared tasks in this domain, such as TREC Legal, FIRE, CLEF, NTCIR, SemEval, and so on.

Topic 6 – The industry perspective: Here we discuss the problems that are particularly important from an industry point of view. We start by briefly covering the three key task-focused areas that legal knowledge workers pursue: finding, analyzing, and making decisions about information. These are critical areas where legal professionals develop refined capabilities in order to address their stakeholders' legal needs. Today, many industries focus on injecting technology (data mining, natural language processing, and machine learning) into the last two tasks: analyzing and deciding. One finds that, within

Table 2. Brief description of some of the resources to be discussed under Topic 5 of the tutorial

Datasets	
Dataset	**Description**
Semantic Segmentation [7]	Corpus of 150 Indian and 50 U.K. Supreme Court Case Documents annotated for the Semantic Segmentation Task
ILSI [28]	Corpus of 65k Indian criminal court case documents for the Legal Statute Identification Task
ILDC [26]	Multiple corpora of Indian Supreme Court case documents for the Court Judgment Prediction and Explanation Task
Summarization [33]	A collection of 3 legal case document summarization datasets
LexGLUE Benchmark [11]	A collection of multiple datasets (mostly EU, UK or US-based) for different legal tasks
Pre-trained Language Models	
Model	**Description**
LegalBERT [10]	Pre-trained Language Model over EU, UK and US legal text
CaseLawBERT [39]	Pre-trained Language Model over US case documents
PoLBERT [20]	Pre-trained Language Model over many types of legal documents
InLegalBERT & InCaseLawBERT [29]	Pre-trained Language Model over Indian court case documents

frequent legal issues, certain patterns repeat and practitioners benefit from seeing such patterns consisting of facts, claims, counter-claims, legal principles applied, analysis and decisions. But statistical discovery of patterns is only possible in large datasets. In today's operational legal settings, one finds ML models, trained on large amounts of data, making predictions in order to assist practitioners with their analysis and decision making. Increasingly such models deploy deep learning to support advanced legal workflows. Some of the challenges that arise around these AI models involve issues such model transparency, explainability, bias, and accountability. We conclude this section by examining some of these evolving challenges.

Topic 7 – Future directions in IR/NLP/ML research in legal domain: We discuss the current problems plaguing the AI & Law scenario, and the future research directions. These include topics like bias and fairness related concerns [17,21], and the need for explainibility in DL models [24,1]. Such issues are of utmost importance to prevent undesirable consequences when such technologies are incorporated into real judicial and legal systems of countries [9,18,38].

Topic 8 – Hands-on coding/demo session: The tutorial concludes with a hands-on demo session to introduce the basics of the coding and modeling strategies to implement and run contemporary models on common legal tasks, such as semantic segmentation [6,7], summarization [5] and court judgment prediction [26]. We also have a look at how citation networks can be used in conjunction to text for tasks such as legal statute identification [28]. This will also cover some existing state-of-the-art resources such as tools and datasets for such tasks.

3   Biography of the tutorial presenters

- **Debasis Ganguly**[8], *Lecturer (Assistant Professor), School of Computing Science, University of Glasgow, Glasgow, Scotland*. Formerly, he was a research staff member at IBM Research Europe, Dublin, Ireland. Generally speaking, his research activities span topics on IR and NLP. More specifically, he is interested in semantic search, neural retrieval models, explainable search and recommendation, fair and trustworthy search, and privacy preserving AI. Apart from this, he is interested in automatically constructing knowledge bases from legal documents for structured and explainable search. He is a part of the organization committee of the Symposium on Artificial Intelligence and Law (SAIL).

- **Jack G. Conrad**[9], *Director of Applied Research, 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, ML and textual data processing. 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.

- **Kripabandhu Ghosh**[10], *Assistant Professor, Department of Computational & Data Sciences, IISER Kolkata, West Bengal, India*. He completed his Ph.D. from Indian Statistical Institute, Kolkata on "Information Retrieval in the Legal Domain". He has been working on AI-Law topics for the last 13 years. His papers have been awarded at the two international conferences that are widely considered to be the most reputed venues in AI and Law – Best Paper award at JURIX 2019 and Best Student Paper award at ICAIL 2021. He has also published papers in AI-Law in reputed journals such as AI and Law, Springer (the most reputed AI-Law journal) and conferences such as SIGIR, CIKM, ACL, ECIR etc. In addition, he has organized several AI-Law events, including a workshop with an international conference (ACM CIKM), shared tasks, and international symposiums such as SAIL'21 [11] and SAIL'22, [12] that hosted talks by reputed researchers in the domain. He has recently been inducted to the Editorial Board of the AI and Law journal, Springer (the most reputed AI-Law journal).

- **Saptarshi Ghosh**[13], *Assistant Professor, Department of Computer Science & Engineering, IIT Kharagpur, West Bengal, India*. His research interests include Legal analytics, Social media analytics, and Algorithmic bias and fairness. 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*

---

[8] https://gdebasis.github.io/   (contact   person)   Email: debasis.ganguly@glasgow.ac.uk

[9] http://www.conradweb.org/~jackg/

[10] https://www.iiserkol.ac.in/web/en/people/faculty/cds/kripaghosh

[11] https://sites.google.com/view/sail-2021/

[12] https://sites.google.com/view/sail-2022/

[13] http://cse.iitkgp.ac.in/~saptarshi

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' (Springer).

– **Pawan Goyal**[14], *Associate Professor, Deptt. of Computer Science & Engineering, IIT Kharagpur, West Bengal, India*. He received his B. Tech. degree in Electrical Engineering from IIT Kanpur in 2007 and his PhD degree from University of Ulster, UK in 2011. He was then a post doctoral fellow at INRIA Paris Rocquencourt. His research interests include Natural Language Understanding, Information Retrieval and Sanskrit Computational Linguistics. He has published his research work in various top conferences including ACL, EMNLP, NAACL, KDD, SIGIR, WSDM, AAAI, IJCAI, as well as journals such as Computational Linguistics, ACM and IEEE Transactions. He was the recipient of Google India AI/ML research awards 2020 and INAE Young Engineers Award 2020.

– **Paheli Bhattacharya**[15], *PhD, Dept. of Computer Science & Engineering, IIT Kharagpur, West Bengal, India*. Her PhD thesis was based on developing text and graph based methods for mining legal court case documents. She has received best paper awards from the two most prestigious conferences in AI and Law – JURIX 2019 [6] and ICAIL 2021 [8]. She actively co-organized shared tasks at FIRE from 2019-2021 and symposiums – SAIL from 2021-2022 and IGLAIS (an Indo-German initiative) in 2021 – all of which were aimed at fostering research in the application of AI techniques in the legal domain. During her MS (by Research) from the same department, she worked on cross-lingual IR on Indian languages. She has publications at SIGIR [3], ECIR [5], AACL-IJCNLP [33], IPM [4], AI & Law [7] and TALLIP. Her research interests are in the area of natural language processing, deep learning and information retrieval.

– **Shubham Kumar Nigam**[16], *Senior Research Fellow, Department of Computer Science &, Engineering, IIT Kanpur, Uttar Pradesh, India*. His research interests include legal data analytics, applications of NLP and IR in the legal domain. Before joining IIT Kanpur, he worked as a project assistant at the Aeronautical Development Agency (ADA), Ministry of Defence, India. He is on the Symposium on Artificial Intelligence and Law (SAIL) organization committee. He actively participates in shared tasks like competitions on legal information extraction and entailment (COLIEE) and semantic evaluation (SemEval). His work has been presented at conferences such as ACL 2021, JURISIN 2022, SemEval 2022, and NLLP 2022.

– **Shounak Paul**[17], *Senior Research Fellow, Department of Computer Science &, Engineering, IIT Kharagpur, West Bengal, India*. His research interests mainly include legal data analytics and applications of NLP in the legal domain. His works on AI & Law for the Indian scenario have been published in premier conferences and journals such as: semantic segmentation [6,7] (JURIX 2019, best paper award; AI & Law Journal 2021), charge identification [27] (COLING 2020) and legal statute identification using citation networks [28] (AAAI 2022).

References

1. Alammar, J.: Ecco: an open source library for the explainability of transformer language models. In: Proc. of ACL-IJCNLP (2021)

---

[14] http://cse.iitkgp.ac.in/~pawang
[15] https://sites.google.com/site/pahelibh/
[16] https://sites.google.com/view/shubhamkumarnigam
[17] https://sites.google.com/view/shounakpaul95

2. Bench-Capon, T., Araszkiewicz, M., Ashley, K., Atkinson, K., Bex, F., Borges, F., Bourcier, D., Bourgine, P., Conrad, J.G., Francesconi, E., et al.: A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. AI & Law (2012)

3. Bhattacharya, P., Ghosh, K., Pal, A., Ghosh, S.: Hier-SPCNet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. Proc. of SIGIR (2020)

4. Bhattacharya, P., Ghosh, K., Pal, A., Ghosh, S.: Legal case document similarity: You need both network and text. Information Processing & Management (2022)

5. Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S.: A comparative study of summarization algorithms applied to legal case judgments. In: Proc. of ECIR (2019)

6. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: Identification of rhetorical roles of sentences in Indian legal judgments. In: Proc. of JURIX (2019)

7. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. AI & Law (2021)

8. Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K., Ghosh, S.: Incorporating domain knowledge for extractive summarization of legal case documents. In: Proc. of ICAIL (2021)

9. Branting, K., Weiss, B., Brown, B., Pfeifer, C., Chakraborty, A., Ferro, L., Pfaff, M., Yeh, A.: Semi-supervised methods for explainable legal prediction. In: Proc. of ICAIL (2019)

10. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Proc. of EMNLP (2020)

11. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., Aletras, N.: LexGLUE: A benchmark dataset for legal language understanding in English. In: Proc. of ACL (2022)

12. Conrad, J.G., Al-Kofahi, K.: Scenario Analytics: Analyzing jury verdicts to evaluate legal case outcomes. In: Proc. of ICAIL (2017)

13. Conrad, J.G., Zeleznikow, J.: The significance of evaluation in AI and Law: A case study re-examining ICAIL proceedings. In: Proc. of ICAIL (2013)

14. Conrad, J.G., Zeleznikow, J.: The role of evaluation in AI and Law: An examination of its different forms in the AI and Law Journal. In: Proc. of ICAIL (2015)

15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of NAACL (2019)

16. Diao, S., Xu, R., Su, H., Jiang, Y., Song, Y., Zhang, T.: Taming pre-trained language models with n-gram representations for low-resource domain adaptation. In: Proc. of ACL-IJCNLP (2021)

17. Garrido-Muñoz , I., Montejo-Ráez , A., Martínez-Santiago , F., Ureña-López , L.A.: A survey on bias in deep NLP. Applied Sciences (2021)

18. Górski, Ł., Ramakrishna, S.: Explainable artificial intelligence, lawyer's perspective. In: Proc. of ICAIL (2021)

19. Governatori, G., Bench-Capon, T., Verheij, B., Araszkiewicz, M., Francesconi, E., Grabmair, M.: Thirty years of Artificial Intelligence and Law: the first decade. AI & Law (2022)

20. Henderson, P., Krass, M.S., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D., Ho, D.E.: Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. arXiv (2022)

21. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. arXiv (2020)

22. Kann, K., Cho, K., Bowman, S.R.: Towards realistic practices in low-resource natural language processing: The development set. arXiv (2019)

23. Leitner, E., Rehm, G., Moreno-Schneider, J.: A Dataset of German Legal Documents for Named Entity Recognition. In: Proc. of the Language Resources and Evaluation Conference (LREC) (2020)

24. Lertvittayakumjorn, P., Toni, F.: Explanation-Based Human Debugging of NLP Models: A Survey. Transactions of the Association for Computational Linguistics (2021)
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv (2019)
26. Malik, V., Sanjay, R., Nigam, S.K., Ghosh, K., Guha, S.K., Bhattacharya, A., Modi, A.: ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In: Proc. of ACL-IJCNLP (2021)
27. Paul, S., Goyal, P., Ghosh, S.: Automatic charge identification from facts: A few sentence-level charge annotations is all you need. In: Proc. of COLING (2020)
28. Paul, S., Goyal, P., Ghosh, S.: Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from Indian legal documents. Proc. of AAAI (2022)
29. Paul, S., Mandal, A., Goyal, P., Ghosh, S.: Pre-training transformers on Indian legal text. arXiv (2022)
30. Şahin, G.G.: To augment or not to augment? a comparative study on text augmentation techniques for low-resource NLP. Computational Linguistics (2022)
31. Sartor, G., Araszkiewicz, M., Atkinson, K., Bex, F., van Engers, T., Francesconi, E., Prakken, H., Sileno, G., Schilder, F., Wyner, A., et al.: Thirty years of Artificial Intelligence and Law: the second decade. AI & Law (2022)
32. Savelka, J., Walker, V., Grabmair, M., Ashley, K.: Sentence boundary detection in adjudicatory decisions in the united states. TAL (2017)
33. Shukla, A., Bhattacharya, P., Poddar, S., Mukherjee, R., Ghosh, K., Goyal, P., Ghosh, S.: Legal case document summarization: Extractive and abstractive methods and their evaluation. In: Proc. of AACL (2022)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proc. of NeurIPS (2017)
35. Villata, S., Araszkiewicz, M., Ashley, K., Bench-Capon, T., Branting, L.K., Conrad, J.G., Wyner, A.: Thirty years of Artificial Intelligence and Law: the third decade. AI & Law (2022)
36. Xiao, C., Hu, X., Liu, Z., Tu, C., Sun, M.: Lawformer: A pre-trained language model for chinese legal long documents. AI Open (2021)
37. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Proc. of NeurIPS (2019)
38. Yu, W., Sun, Z., Xu, J., Dong, Z., Chen, X., Xu, H., Wen, J.R.: Explainable legal case matching via inverse optimal transport-based rationale extraction. In: Proc. of SIGIR (2022)
39. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E.: When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In: Proc. of ICAIL (2021)