IJCNLP-AACL 2025 Tutorial Proposal

📌**Title**

Trustworthy Legal Text Processing with LLMs: *Retrieval, Rhetorical Roles, Summarization, and Trustworthy Generation*

📝 **Abstract**

This half-day tutorial provides a comprehensive overview of **Legal Natural Language Processing (NLP) with LLM** for participants with a basic understanding of Computational Linguistics or NLP concepts. We introduce how NLP can help analyze and manage legal text by covering five key topics: legal text analysis with LLM insights, legal text retrieval, rhetorical role identification, legal text summarization, and addressing bias and hallucination in legal tasks. Our goals are to explain why these tasks matter for researchers in the legal domain, describe the challenges and open problems, and outline current solutions. This proposed tutorial blends lectures, live examples, and Q&A to help researchers and students see how language technology and LLMs can make legal information more understandable and efficient.

📖 **Introduction**

Legal NLP is about applying language technology to law using computers to process and understand legal documents such as cases, statutes, and contracts. This is timely because modern law generates massive amounts of text, and new AI models have brought tools into the public eye. For example, legal professionals often spend hours or days reading cases and statutes, a burden that technology could help reduce. Retrieving relevant precedents already "comprises much of a lawyer's time." At the same time, accessible AI tools create excitement and raise issues, promising to make legal knowledge more democratic and faster. Yet, they risk mistakes or unfairness if not handled carefully.

In this context, our proposed tutorial describes five subtopics of legal NLP. We explain each topic's importance and challenges. For example, legal documents are often very long and complex. So, searching and summarizing them is complex. We illustrate how NLP models can extract useful information from large corpora of legal texts, annotate case documents with rhetorical roles, and produce clean, concise, and readable summaries. We also highlight the use of LLMs and challenges in addressing historical biases and hallucinated or fabricated legal content.

By the end, participants will understand broader societal and legal implications of these technologies and gain a comprehensive view of the latest state-of-the-art developments in the field.

🎯 Target Audience

This tutorial helps researchers and professionals in the legal and AI domains. Attendees could include legal domain researchers, LLM researchers, legal experts, policymakers, and scholars interested in how AI can support legal practice and research.

We expect roughly 60 attendees, given the growing interest in legal AI and the broad appeal to both legal and AI-interested communities. We will explain core ideas through real-world examples, making the content accessible with a balanced blend of theory and illustrative demos.

📑 Outline

The tutorial is organized into the following segments.

- **Introduction to Legal NLP and LLMs (~15 Slides—30 minutes):** We explain legal NLP and why it matters. We cover key motivations: the vast amount of legal text worldwide, the costs of manual review, and recent advances in AI. We mention concrete use cases and outline examples of NLP in law highlighting the challenges.

- **Legal Text Analysis & Extraction (~15 Slides—30 minutes)**: This section covers how NLP can analyze legal text by rule-based methods, machine learning, and deep learning approaches (NER, relation extraction, and document classification). We demonstrate the extraction of legal precedents and citations by discussing its usefulness and the challenges of legal language, such as length and dense jargon.

- **Legal Text Retrieval (~15 Slides—30 minutes)**: We explain how NLP helps search legal databases by combining LLMs with external knowledge for accuracy and trust. We focus on statute retrieval, finding laws and regulations, and introduce RAG, which blends document search with answer generation. A key challenge is that legal queries often require matching exact fact patterns, not just keywords.

- **Break for 30 Minutes**

- **Rhetorical Role Identification (~10 slides—20 minutes):** Many legal documents, especially court opinions, have an implicit structure of sections like Facts, Arguments, Rules, and Decision. Identifying these rhetorical roles helps readers and AI systems alike. We explain this concept: for example, one task labels each sentence of a case as "Facts", "Lower Court Ruling", "Argument", "Precedent", "Decision", etc. We show why knowing where the facts end and the judgment begins can speed up research or summarization. We will present a simple example of a case excerpt and manually point out the roles, then discuss how an NLP system might do it.

- **Legal Text Summarization (~15 Slides—30 minutes)**: We introduce summarization, which is condensing long documents into shorter, simpler versions. In law, summarizers can turn a long opinion or contract into a digest or bullet points. We explain the benefit: legal professionals often spend days reading documents, so summaries can help to

condense lengthy legal documents into concise summaries and save both time and costs. We cover the two main styles of summarization in simple terms: extractive—picking key sentences vs. abstractive—rewriting in new words. We also discuss limitations: it's difficult to capture legal nuance, and poor summaries can omit essential details or introduce errors.

- **Bias and Hallucination in Legal NLP (~15 Slides—30 minutes):** We discuss two significant ethical challenges. First, **bias**—AI trained on historical legal data can inadvertently learn past prejudices or systematically misinterpret how laws apply to different groups. We discuss the importance of fairness, where AI should not disadvantage marginalized groups, and the need for careful design and evaluation. Second, **hallucination**—AI can confidently produce false information. In legal contexts, it is dangerous. We explain the risks and emphasize best practices.

- **Q&A and Discussion (10 minutes):** We conclude by reviewing key points, answering remaining questions, and discussing future directions. We invite participants to consider how legal NLP might affect their work or society and encourage critical discussion.

🌐 **Diversity Considerations**

We believe this tutorial explains fairness and inclusion in several ways:

1. **Broad Accessibility**: By focusing on an introductory presentation style, we make AI knowledge accessible to people doing research in NLP tasks like summarization, extraction, retrieval, and LLMs, as well as legal experts. This helps diversify the community that is engaging with NLP technology.

2. **Relevance to Underrepresented Groups**: Legal NLP can directly impact fairness in society. Law and technology have both historically underrepresented certain voices. Discussing bias, data equity, and inclusive design raises awareness. We will highlight examples where attention to diverse data is essential, thus encouraging participants to consider varied perspectives in legal AI.

📚 Reading List

To accommodate different levels of background, we recommend the following resources:

Introductory Materials (for general understanding):

- Kevin D. Ashley (2017). "Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age." (A non-technical book showing how AI methods apply in law.)

Recommended Reading (before attending):

- Ariai, F. and Demartini, G. (2024). "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges." (A recent survey summarizing legal NLP tasks and challenges.)

- M. P. Prajwal and Anand Kumar Madasamy (2022). "Legal Text Analysis Using Pre-trained Transformers." (A study on legal text analysis)

- Locke, D. and Zucccon, G. (2022). "Case Law Retrieval: Accomplishments, Problems, Methods and Evaluations." (A review of legal search methods.)

- Muhammed, A., Muslihuddeen, H., Sankar, S., & Anand Kumar, M. (2024). "Impact of Rhetorical Roles in Abstractive Legal Document Summarization." (A study on Rhetorical Roles used in summarization)

- Akter, M., Cano, E., et al. (2025). "A Comprehensive Survey on Legal Summarization: Challenges and Future Directions." (A survey of methods for summarizing legal documents.)

- Bhattacharya, P., Dash, P., et al. (2023). "DeepRHOLE: Deep Learning for Rhetorical Role Labeling of Legal Case Sentences." (A paper on identifying sections in legal judgments.)

Advanced/Supplementary Resources:

- Zhong, H., et al. (2020). "How does NLP benefit legal system: A summary of legal artificial intelligence" (Research on AI predicting legal judgments.)

- Charlotin, D. (maintained online). "AI Hallucination Cases" (a continually updated list of court cases discussing AI errors in law).

- Sindhu, P., Gupta, D., & Meghana, S. (2023). "NITK_LEGAL at SemEval-2023 Task 6: A Hierarchical based system for identification of Rhetorical Roles in legal judgements." *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

- Muhammed, A., Muslihuddeen, H., Sankar, S., & Anand Kumar, M. (2024). "SCaLAR NITK at SemEval-2024 Task 5: Towards Unsupervised Question Answering System with Multi-level Summarization for Legal Text." *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

- Anu Thomas, Sangeetha Sivanesan, An adaptable, high-performance relation extraction system for complex sentences,Knowledge-Based Systems, Elsevier, Volume 251,2022,

---

👥 **Presenters**

1. **Dr. M. Anand Kumar** is an Associate Professor in the Department of Information Technology at the National Institute of Technology Karnataka (NITK), with over 14 years of academic experience. His research spans Natural Language Processing (NLP), information retrieval, text analytics, machine translation for Indian languages, legal document summarization, explainable AI, and applied machine learning for social and legal domains. He is the principal investigator of a **ANRF-SERB-CRG (2024-2027) project on "A Deep explainable framework for semantically similar document retrieval and summarization of legal text."** He has led multiple funded projects and consultancy assignments, including language technology development for Tamil and Dravidian languages. He has over 200 Scopus-indexed publications, 2,200+ citations, and has organized more than nine international shared tasks in Indian languages. He has delivered tutorials and lectures to diverse audiences, including legal scholars, social scientists, and government professionals. He is experienced in making NLP concepts accessible to technical and non-technical participants. He has taught and mentored undergraduate, postgraduate, and doctoral students from multidisciplinary backgrounds in NLP, with two doctoral students having successfully graduated.

2. **Dr. (Mrs.) S. Sangeetha** is an Associate Professor in the Department of Computer Applications at the National Institute of Technology, Tiruchirappalli. She specializes in Natural Language Processing (NLP) and Information Extraction. She holds Ph.D. from National Institute of Technology, Tiruchirappalli in the broad area of Information Extraction.She actively contributes to academic governance through various committee memberships and has guided numerous postgraduate research projects. Dr. Sangeetha has received several accolades, including the Best Performer Award and Best Paper Awards at national conferences. Her research focuses on NLP applications and she has delivered invited talks and coordinated workshops in her field including **Legal Artificial Intelligence**. She also supervised a Ph.D thesis titled **Intelligent and Adaptive Information Extraction from Indian E-Judgments Towards Constructing Knowledge Graph in Judicial Domain.** She is a life member of the ISTE and a member of the Association of Computational Linguistics.

3. **Dr. Manikandan Ravikiran** is a Lead Research Scientist at ThoughtWorks AI Research Lab (Global Team), where he focuses on enhancing Large Language Model (LLM) adoption through evaluation frameworks targeting completeness, explanation fairness, and decoding optimization for multiple domains. Parallel to his industry role, he is a Ph.D. student at the Indian Institute of Technology, Mandi, advised by Prof. Rohit Saluja and Prof. Arnav Bhavsar, researching rare and impactful problems in AI for the education and legal domains. His research interests span fairness in LLMs, explanation sufficiency, multilingual and educational NLP, **legal AI**, low-resource modeling, and deep model compression, with a special emphasis on developing evaluation methods and scalable attention mechanisms for complex reasoning tasks in real-world AI systems. He has served on the program committees of ACL, NAACL, EMNLP, COLING, AACL, LREC,

NeurIPS, IEEE ICDDS, ACM ICMR, Springer Language Resources and Evaluation (LRE), ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Journal of Experimental and Theoretical Artificial Intelligence (JETAI), and Elsevier Engineering Applications of Artificial Intelligence (EAAI). He received the ACL 2023 Outstanding Reviewer Award and has been a visiting researcher at the Department of CSE, IIT Kanpur, and NII Japan. He has co-organized multiple workshops and tutorials, including the Workshop on Cross-Modal Learning (WCRML 2019) at ACM ICMR, the Second and Third Workshops on Speech and Language Technologies for Dravidian Languages at EACL (2022, 2023), and the IEEE Bigdata CONSTRAINT Workshop (2023, 2024, 2025).

4. **Anjali R** is currently pursuing her full-time PhD under the guidance of Dr. Anand Kumar M. in Information Technology at the National Institute of Technology Karnataka, Surathkal, with research interests spanning Natural Language Processing (NLP), computer vision and deep learning. She has delivered invited talks on generative AI and the fundamentals of NLP, contributing to academic and research communities through knowledge sharing.

ℹ️ Other Information

Expected Number of Attendees: We estimate about 60 participants, based on interest in legal tech and previous similar tutorials. The IJCNLP-AACL community is growing, and legal technology appeals to both language researchers and domain experts.

Equipment/Setup: We will use slides and require a projector and sound for videos. An Internet connection is needed for live demos. Presenters will bring a laptop with pre-configured demo environments. No special hardware is required; all tools are software-based.

• **Room Logistics**: A lecture room with seating for up to 60, plus a whiteboard.

• **Additional Notes**: As a half-day event, we will include a 15-minute break. We encourage attendees to bring their questions.

⚖️ Ethics Statement

Use of Legal Data and Privacy: Legal documents often contain sensitive information (names, personal data, privileged details). Any NLP or AI applied to such texts must respect confidentiality. We will stress that many advanced systems are trained on publicly available cases but caution that private documents should not be used without consent.