

# Tutorial on Legal Text Analytics

**Vasudha Bhatnagar, Vikas Kumar**

Dept of Computer Science, Faculty of Mathematical Sciences, University of Delhi

**Balaji Ganesan**

IBM Research, India

**Muhammad Abulaish**

Dept of Computer Science, South Asian University

**Parikshet Sirohi, Ashwini Siwal**

Law Center - II, Faculty of Law, University of Delhi

vbhatnagar@cs.du.ac.in

## Abstract

Legal research is a difficult and crucial task in the practice of law. It requires intense human effort and intellectual prudence to research a legal case and prepare arguments. By extracting targeted information, NLP-powered applications have the potential to significantly speed up this process. In this tutorial, we'll outline the most recent findings in legal text analytics with an emphasis on Indian judicial records.

## 1 Introduction

The use of information technology can significantly reduce the number of court cases that are pending in India. Reducing court backlogs is a necessary first step toward increasing the accessibility of justice for the common person and the transparency of the judicial system. One of the successful projects in the nation built on free and open-source software is the eCourts mission mode project, which was launched in 2005. Initiatives in this area are anticipated to expand access to the legal profession's expertise and generate significant societal benefits.

Natural language processing (NLP) has made revolutionary strides in recent years, igniting the burgeoning field of Legal Artificial Intelligence (Legal AI). Legal AI aims to use NLP technology to create tools for the automatic extraction of desired information to assist stakeholders in the legal domain. The objectives and future prospects of NLP research in Legal AI are described in (Zhong et al., 2020).

With an emphasis on Indian legal text, we seek to deliver state-of-the-art legal text analytics in this tutorial. We'll discuss large language models (LLMs), which are frequently employed for summarizing legal documents, question-answering,

Info	Details
<b>Title</b>	Tutorial on Legal Text Analytics
<b>Scope</b>	We'll cover NLP topics relevant for Legal Text Analytics.
<b>Structure</b>	Introduction - 20 mins Legal KG - 40 mins Question Answering on Indian Judgements - 40 mins Legal Text Classification - 40 mins Judgement Summarization - 40 mins Total - 180 mins
<b>Presenters</b>	Balaji Ganesan Muhammad Abulaish Vikas Kumar Vasudha Bhatnagar
<b>Citations</b>	Please see section 2.
<b>Audience</b>	Academics and industry practitioners working in NLP.
<b>Background</b>	Familiarity with Machine Learning is useful but not required.

Table 1: Essential Information on the proposed tutorial

text classification, and creating knowledge graphs from legal documents, as well as NLP techniques for analyzing legal documents that are supported by neural networks. We'll present the four aforementioned topics in the context of Indian court judgments, which are the most common documents available in Indian legal proceedings.

We will highlight the main ongoing initiatives in this field as well as the research efforts done by the Indian NLP community. We anticipate that the seminar will inspire young scientists, particularly

those engaged in NLP and text mining research.

## 2 Tutorial Content

Please refer to Table 1 for the tutorial structure.

### Legal Knowledge Graph

Automatic Knowledge Graph Construction (AKBC) has been popularized since the Knowledge Base Population track (Ji et al., 2010) organized by TAC. Recent works have focused on personal knowledge base population among others. However domain specific knowledge graphs (Abu-Salih, 2021) remain an ongoing research problem. In (Dhani et al., 2021), we discussed creating a legal knowledge graph using judgements and related documents from Indian courts. We'll present the state of the art in legal knowledge graphs.

### Question Answering on Indian Judgements

Although question answering is a well-studied problem (Pramanik et al., 2021), QA systems in the legal domain are not commonly available. Automatic question answering systems not only provide consultancy to litigants who are typically chartering unfamiliar grounds of the legal domain, but are equally beneficial to legal professionals. In (Abulaish and Dey, 2007) and (Ganesan et al., 2020), we had presented question answering in biology and compliance domains. In this talk, we'll present the latest works in question answering on legal documents, especially on Indian court judgements.

### Legal Text Classification

Text classification is again a well-studied problem (Kumar, 2019). Legal text classification too has been discussed in (Sulea et al., 2017; Wei et al., 2018; Shaheen et al., 2020). The presence of terms unique to legal texts, many of them latin words and phrases, the difference in legal processes in different countries makes this task harder on legal text (Chalkidis et al., 2019). Especially, in the Indian context, text classification requires huge amounts of manually annotated data (Dhani et al., 2021). We'll present recent efforts in this space.

### Judgment Summarization

Automatic summarization of judgements, and preparation of *Headnotes* (highlighting the point-of-law) help law professionals to locate discussion of a legal issue in lengthy judgements. There have

been works in different countries that have addressed this including SALOMAN (Uyttendaele et al., 1998) in Belgium, Letsum (Farzindar and Lapalme, 2004) in Canada, Case summarizer (Polsley et al., 2016) in Australia. (Yizhen et al., 2021) summarizes contents of Chinese civil judgments. (Kanapala et al., 2019) presented a survey of works in legal text summarization. In this talk, we'll present these works and discuss Headnotes preparation on Indian court judgements.

## 3 Team

The team has been collaborating for the past year on Masters's projects at University of Delhi.

**Vasudha Bhatnagar** is a *Professor* in the Department of Computer Science, University of Delhi. She has extensive experience in data mining, machine learning, network science and text analytics. She obtained her Ph.D. from Jamia Millia Islamia, New Delhi.

**Muhammad Abulaish** is a *Professor and Chairperson* of the Department of Computer Science, South Asian University, New Delhi. His research interests are in data analytics and machine learning, with applications in text mining, social network analysis, and biomedical informatics among others. Abulaish received his Ph.D. from IIT Delhi.

**Vikas Kumar** is an *Assistant Professor* in the Department of Computer Science, University of Delhi. His primary research interests are data mining and machine learning. Gold Medalist in MCA from Pondicherry University, he obtained Ph.D. degree from the University of Hyderabad.

**Balaji Ganesan** is a *Senior Research Engineer* at IBM Research, India. He currently works on search, knowledge graphs and explainability. Balaji was previously at Yahoo, and a number of startups. Balaji has a Master's degree in Computer Science from the University of Arizona.

**Parishet Sirohi** is an *Assistant Professor* in the Faculty of Law at the University of Delhi. He has over sixteen years of experience in the areas of industry, litigation and academics. He is currently teaching the subjects of Intellectual Property Rights and Constitutional Law. He was awarded the Ph.D. degree by the Faculty of Law, Jamia Millia Islamia. **Ashwini Siwal** is an *Assistant Professor* in Intellectual Property Law where he specializes in intellectual property Laws, data security laws, and science, technology and law. He obtained his Ph.D. from Jamia Millia Islamia, New Delhi.

## References

- Bilal Abu-Salih. 2021. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185:103076.
- Muhammad Abulaish and Lipika Dey. 2007. Biological relation extraction and query answering from medline abstracts using ontology-based text mining. *Data & Knowledge Engineering*, 61(2):228–262.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in eu legislation. *arXiv preprint arXiv:1905.10892*.
- Jaspreet Singh Dhani, Ruchika Bhatt, Balaji Ganesan, Parikshet Sirohi, and Vasudha Bhatnagar. 2021. Similar cases recommendation using legal knowledge graphs. *arXiv preprint arXiv:2107.04771*.
- Atefeh Farzindar and Guy Lapalme. 2004. Letsum, a text summarization system in law field. In *a THE FACE OF TEXT conference (Computer Assisted Text Analysis in the Humanities)*, pages 27–36.
- Balaji Ganesan, Avirup Saha, Jaydeep Sen, Mathleen Ahmed Pasha, Sumit Bhatia, and Arvind Agarwal. 2020. Anu question answering system. In *ISWC (Demos/Industry)*.
- Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffith, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third text analysis conference (TAC 2010)*, volume 3, pages 3–3.
- Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.
- Vikas Kumar. 2019. Collaborative filtering and multi-label classification with matrix factorization. *arXiv preprint arXiv:1907.12365*.
- Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.
- Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Uniqorn: unified question answering over rdf knowledge graphs and natural language text. *arXiv preprint arXiv:2108.08614*.
- Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. *arXiv preprint arXiv:2010.12871*.
- Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.
- Caroline Uyttendaele, Marie-Francine Moens, and Jos Dumortier. 1998. Salomon: automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligence and Law*, 6(1):59–79.
- Fusheng Wei, Han Qin, Shi Ye, and Haozhen Zhao. 2018. Empirical study of deep learning for text classification in legal document review. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3317–3320. IEEE.
- Wang Yizhen, Ou Shiyan, and Chen Jinju. 2021. Automatic abstracting civil judgment documents with two-stage procedure. *Data Analysis and Knowledge Discovery*, 5(5):104–114.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.