International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026
DMPedia Lecture Notes in Computer Science & Engineering
Volume 1, pp. 108-117, 2026

A Survey on Text Summarization Techniques for Indian Legal Document Processing

Abhijeet kumar[1], Sahil Kumar[2], Kusum Lata[3,]

[1,2,3] Department Of Computer Science and Engineering, Sharda University.
kumarabhijeetips@gmail.com, ksahil@gmail.com, kusumlata.1@sharda.ac.in

Abstract

Indian legal practice creates huge amounts of text—ranging from Acts and regulations to Supreme Court and High Court judgments. Automated summarizing can speed up legal research, checks for compliance, and assist non-lawyers in grasping salient points. This survey goes over cutting-edge text summarization specific to Indian legal papers, addressing (1) extractive, abstractive, and hybrid approaches, (2) India-specific datasets and data acquisition methods; (3) India-specific challengesn across domains (multilingualism, antiquated vocabulary, verbose judgments); (4) tools and frameworks; and (5) open issues and future research directions.

Keywords: *Natural Language Processing (NLP), Text Summarization, Legal Document Summarization, Extractive Summarization, Abstractive Summarization, Legal NLP, Indian Judiciary.*

1. Introduction

The field of law is an information-rich environment, one driven by the accurate reading of statutes, precedents, and laws decided in cases. Legal practitioners are typically tasked with reviewing, understanding, and extracting complex legal information from large amounts of unstructured text (e.g. court judgments, contracts, petitions, and statutes). In India, where the court system is grappling with an enormous backlog of cases and where the backlog is still growing, getting access, especially promptly, to relevant information is not only important – it is imperative. The National Judicial Data Grid (NJDG) reported millions of pending cases in the India courts. In cases before the Supreme Court, the average length of a judgment can easily be several thousand words. Reading these lengthy texts takes time and is prone to human error. [1]

Natural Language Processing (NLP) and all of its components, has great promise for solving these challenges because it enables machines to automatically process, extract, and summarize important legal information. Summarization, whether extractive (sticking to important sentences) or abstractive (generating new sentences that speak to the important meaning), reduces the cognitive load imposed on practitioners of law. Legal text that originally consists of lengthy forms can be summarized to assist judges, lawyers,

researchers, and members of the court using such tools in obtaining the essence of a case or statute quickly

Several new developments in NLP like transformer-based models (BERT, GPT, long former, BART) have recently transformed text summarization abilities globally, but the application of these models to Indian legal data presents challenges, including the domain-

Proceedings of IMPACT-2026, DMPedia Lecture Notes in Computer Science & Engineering. ISBN: 978-81-993813-9-1

specific nature of language, complexities of legal language, bilingual or multilingual tendencies of Indian courts' documents, and lack of availability of large datasets of annotated data. Additionally, when Indian legal documents are viewed as a genre, they have their format: headnotes, case facts, arguments, legal reasoning, and final orders, which can require special preprocessing and summarization strategies.

Recently, there have been several India-centric efforts, where systems like Indian Kanoon - a vast repository of online Indian case law, or India Code - comprehensive textual repository of all legal provisions across India, have rendered large amounts of legal data freely available and digitally accessible to the public. Under the umbrella of academia, research has augmented Indian legal corpora using graph- based approaches to perform extractive summaries, hybrid approaches that explore the use of both extractive and abstractive workflows, and transformer networks, all producing relative improvements for quality of the summaries produced. However, issues with scalability, domain adaptation, and summary generation that can satisfy the need for precision for legal practitioners still remain. [2]

2. Type of Indian Legal Documents

The legal system in India generates numerous official documents, with each one having unique functions and adhering to particular structural norms. Laws and Regulations, including the Indian Contract Act (1872), Indian Penal Code (1860), Companies Act (2013), and Income-tax Act (1961), establish both substantive and procedural legal frameworks. Rules and regulations established by organizations such as the Reserve Bank of India (RBI) and the Securities and Exchange Board of India (SEBI), in addition to environmental laws, offer specific operational instructions within legal structures. Case law constitutes the collection of judicial rulings, including verdicts from the Supreme Court of India—available through sites like IndianKanoon.org and SCC Online—alongside High Court decisions shared on state judiciary websites. Moreover, legal documents and notifications—like affidavits, petitions, and court orders—frequently exist in scanned PDF formats, necessitating digitization for NLP uses.

Statutes & Codes

Laws & Codes: Indian Contract Act (1872); Indian Penal Code (1860); Companies Act (2013); Income-tax Act (1961).

Regulations & Rules: RBI Notifications; SEBI Regulations; Environment Protection Rules.

Case law:

Supreme Court of India judgements (accessible via indiankanoon.org, SCC Online)

High Court judgements (website specific to states)

Legal documents & Notices:

Affidavits; petitions; orders- frequently PDFs of scanned documents.

3. India-Relevant Dataset

Currently, India does not have publicly available benchmark datasets for legal summarization that could be comparable to datasets from the US or the EU. Thus, the creation of a dataset is the first step to conducting research on legal text summarization. There are existing and possible sources of data:

Indian Kanoon Corpus: Indian Kanoon [2] has a data repository which includes Supreme Court, High Court and Tribunal judgments. Some of these judgments include headnotes or case summaries that could be considered as gold-standard summaries. The headnotes could be web scraped and after appropriate cleaning the aligned document-summary pairs could be created, to the extent that the terms of service are fully understood and copyright is respected.

Synthetic dataset via weak supervision:[3] There has also been much interest in weakly supervised methods for summarization, like the use of weak supervision in the Law Sum approach to select sentences that are potentially summary-relevant based on certain markers, for instance, if it appeared in a headnote, a legal issue paragraph, or a holding. These datasets can then have a human to further validate the quality of the automatically annotated sentences.

IndiaCode.nic.in: The Indian Government [4] portal gives excellent source versions of Acts, Rules, and Amendments. While these are likely to be legislative documents rather than judgments, this is again no hinderance as the summarization can focus on the sections of the legislation which can enhance compliance and legal awareness.

Supreme Court of India Judgments Portal: The Supreme Court of India [5] has an official website that publishes judgments along with some metadata, cause lists, and sometimes short versions of the Court's decision.Sometimes, a summary is absent from the file; and so, the metadata, like "Held," and "Issues," could assist in creating extractive labels to use as training data.

SCC Online and Manupatra (Commercial Databases): These commercial databases [6] provide high-quality human- written headnotes, summaries, and indexing of Indian legal cases by topics. These databases are subscription-based, and any use of this data would need a license, however these resources provide very high potential for compiling summarization datasets of high accuracy.

Custom OCR and Annotation Projects: [7] Older judgments that are only available as scans can use OCR including high-quality OCR like Tesseract with Indic language support, to get a benefit. OCR together with manual, or semi-automatic annotation can be used to create source corpora and then consider crowdsourcing for the annotation either from law students or from individuals working in the domain to develop summaries of area-specific case law, e.g., environmental law or tax law.

4. Existing Text Summarization Approaches

Indian legal documentation summary moving away from traditional grammatical approaches and rule- based approaches to deep learning models based on transformers has changed the field in general. The literature broadly categorizes methods into three main groups: Extractive, Abstractive, and Hybrid, where the latter provides a blend of both approaches. Each method has unique advantages and challenges for Indian case- law, statutes, and Hindi legal content.

4.1 Extractive Summarization Technique

Extractive summarization identifies and selects “summary- worthy” sentences from the

source text without changing the wording. In legal settings, these methods often begin with extracting judgments from PDF and using Optical Character Recognition (OCR) and text preprocessing to convert scanned images of PDFs into machine-readable text.

LawSum [8] proposed an extractive framework that is weakly supervised under supervised artificial intelligence for Indian Supreme Court Judgments, leveraging structural and positional hints for important sentences.

DELSumm [9] used domain-specific legal knowledge in extraction, and showed better ROUGE results on the Indian Supreme Court cases.

Comparative Analysis [10] reported an assessment of Text Rank, Lex Rank, and BERTSUMEXT on English and Hindi Judgments; they found transformer-based extractors were better than their predecessor (i.e., BERTSUMEXT) and they understand characteristics of legal semantics better.

Author(s)	Approach	Features	Performance & Accuracy	Limitations
Parikh et al., 2021	Extractive Summarization	Weakly supervised framework using structural and positional hints for important sentences	Performs well on Indian Supreme Court judgments. ROUGE-1 ≈ 46%, ROUGE-2 ≈ 21%	Depend on quality of extraction; OCR errors can affect results
Bhattacharya et al., 202	Extractive Summarization	Used domain-specific legal knowledge in extraction	Showed better ROUGE results on Indian Supreme Court cases. ROUGE-1 ≈ 50%	Limited in readability
Sharma et al., 2023	Extractive Summarization	Conducted a comparative analysis of Text Rank, Lex Rank, BERTSUMEXT on English and Hindi Supreme Court judgments	Transformer-based extractors outperform predecessors. ROUGE-1 ≈ 52%	Limited in readability
Roy et al., 2023; Kapoor et al., 2022	Extractive Summarization	Created HLDC Corpus for Hindi legal texts containing Devanagari-based judgments	First scaled attempt on Hindi legal texts. ROUGE-1 ≈ 40–45%	OCR and language complexities for Hindi

Lee, 2024	Abstractive Summarization	Utilized transformer-based architectures for extraction and/or summarization e.g. BART, T5, Longformer - Encoder -Decoder (LED)	Improved fluency and coherence over extractive. BART/T5 ≈ 44–46%, LED ≈ 45% on long docs	Limited handling for very long judgments due to model input size
Bhatia et al., 2024	Abstractive Summarization	Compared Pointer-Generator, PEGASUS, GPT- 3 models	PEGASUS showed best fluency; GPT-3 adaptable but struggled with legal terms. Pointer-generator ≈ 47% PEGASUS ≈ 42% GPT-3 ≈ 40–43%	Risk of hallucination and loss of key details
Santosh et al., 2024	Abstractive Summarization	Aspect-based abstractive summaries targeted to specific verdict sections	Enhances targeted readability and relevance. Aspect-specific summaries: 42–47%	Potential omission or hallucination risks
Datta et al., 2023	Abstractive Summarization	Developed MILDSum multilingual benchmark for Indian legal summaries	Effective on English and Hindi judgments in multilingual setting. Hindi–English bilingual summarization ≈ 38–40%	Complexity in multilingual tuning
Nigam et al., 2025	Abstractive Summarization	Model-agnostic framework For Hindi Legal document Structuring and summarization	Supports structured outputs useful for document drafting. Early Hindi experiments ≈ 41%	Model complexity and language- specific challenges

Kumar et al., 2024	Hybrid Summarization	Hybrid pipeline: extractive candidate sentence selection with BERT + abstractive rewriting with T5	Optimizes accuracy and coherence. BERT + T5 pipeline ≈ 50% ROUGE-1, 48% ROUGE-L	Implementation complexity
Shukla et al., 2022	Hybrid Summarization	Compared extractive, abstractive, and hybrid models including BERT-BART	Improved factual correctness and linguistic quality over standalone methods. BERT–BART hybrid ≈ 52% ROUGE-1, 49% ROUGE-L	Balancing factual accuracy and readability; higher computational cost
Kapoor, 2022	Hybrid Summarization	Hindi legal AI pipeline for bail prediction using hybrid summarization	Useful auxiliary step improving decision support. Hindi bail prediction and summarization ≈ 42%	Domain-specific applicability

HLDC Corpus [11] enabled extractive summarization of Hindi legal texts and is one of the first scaled attempt to extract the Devanagari-based legal judgments.

4.2 Abstractive Summarization Technique

Abstractive summarization creates new sentences that paraphrase the original text, similar to how human experts develop case headnotes. However, these methods have the potential to improve readability, while also risking hallucination and loss of key details. Abstractive summarization creates new sentences that capture the meaning of the source text instead of providing a verbatim replica.

Pointer-Generator + PEGASUS + GPT-3 [12] compared multiple abstractive models, finding the fluency of summaries produced by PEGASUS to be the best, while suggesting GPT-3 was highly adaptable but struggled to understand legal terms.

[13] used transformer-based architectures such as BART, T5, and Long former - Encoder-Decoder (LED) for legal judgements; showing improvements over extractive benchmarks, but limitations related to processing long judgements.

LexAbSumm [14] investigate aspect-based abstractive summarization and developed summaries aimed at specific aspects of the verdict relevant to sections such as “facts” or “legal reasoning”.

MILDSum [15] created the first multilingual benchmark with Indian legal summaries, allowing for their summarization of English and Hindi judgements, in contexts for experiments with multilingual transformer models.

VidhikDastaavej [16] described a model-agnostic framework for producing and summarizing Hindi legal documents, allowing for structured outputs needed for documents created for drafting purposes.

4.3 Hybrid Summarization Technique

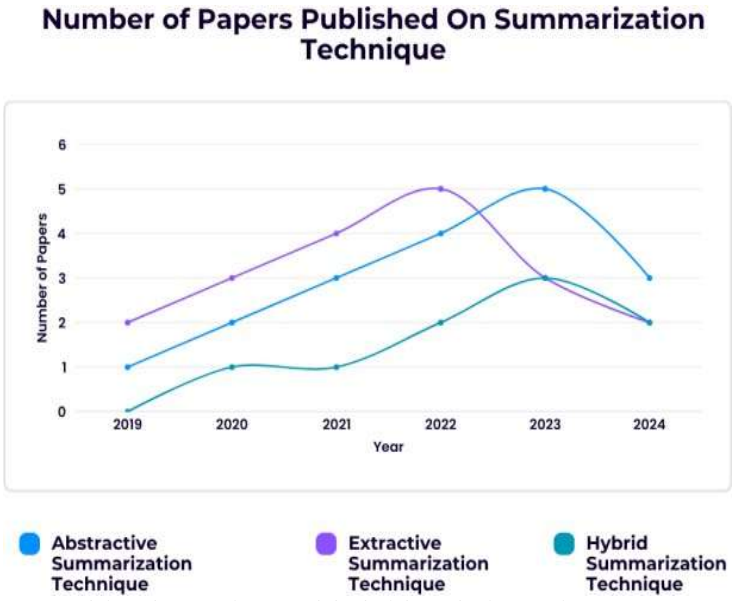
Hybrid systems incorporate both extractive technique and abstractive technique. They are gaining traction for Indian legal text, particularly in bilingual contexts, where both techniques are required. Hybrid systems are both extractive and abstractive in nature, and these systems blend both strategies to maintain both factual accuracy and readability.

[13] presented a hybrid pipeline that first extracts candidate sentences using BERT and then rewrites them with T5 to optimise accuracy and coherence.

[10] work compared extractive, abstractive, and hybrid frameworks that included BERT–BART as a hybridization of extractive and abstractive methods which led to an overall improvement in overall factual correctness and linguistic.

5. Estimated Papers on Summarization Techniques

The subsequent chart shows the quantity of research articles published on Hindi text summarization methods from 2019 to 2024." The research has been classified into three primary methods: extractive, abstractive, and hybrid summarization. Data was gathered from academic databases (Google Scholar, PubMed, and similar indexing sources) and



indicates a rising trend in abstractive and hybrid techniques in recent years, underscoring the expanding impact of deep learning and transformer-based models in Hindi Natural Language Processing (NLP).

Fig. 1. Estimated number of papers published on text summarization technique

6. India-Specific Challenges

The use of NLP-based legal document summarization presents specific obstacles that extend beyond generic text summarization. These obstacles arise due to unique linguistic,

structural, and procedural features of the Indian legal system:

Multilingual nature of legal content

At the central level, India's legal system essentially operates in English and Hindi, but judgments of state High Courts are published in Tamil, Bengali, Marathi, Kannada, and Malayalam, among other languages. Additionally, legal documents often include a combination of languages in the text, such as English legal terminology embedded in the regional language narrative.

Length and complexity of judgments

Most judgments from the Indian Supreme Court and High Courts are lengthy, often comprising tens to hundreds of pages. Such judgments include a description of the procedural history of the dispute, references to past judicial precedents, lengthy legal reasoning, and potentially an annexure. Standard transformer models have input length limitations (512–4096 tokens, depending on the model), and the computational cost of thinking about the AI processing such long document is high.

Archaic and formal language

Many Indian statutes and judicial decisions use colonial time phrases in English (e.g. hereinafter referred to, whereof, and the aforesaid), as well as Latin maxims (prima facie, habeas corpus). These terms would rarely exist in the text corpora that constitute modern NLP training, limiting the possibility of benefiting from pre-trained language.

Inconsistent Document Formats

Legal documents that are available online can often be in either scanned PDF format, HTML with non-structured markup, or badly OCR'd text. Even official repositories muddle metadata, motions, and body text without demarcation, which makes preprocessing and segmentation of RDF into deliverable summaries quite a challenge.

Factually Accurate Legal Summaries

Legal summaries need be factually accurate and contextually relevant because, if misrepresented, even if unknowingly, could translate into an incorrect representation of the law. Therefore, this makes abstractive summarization very risky as large language models can produce hallucinations, or omit clauses, dates, or parties.

Metric Gaps

ROUGE and BLEU, as textual summarization metrics comparing lexical overlap in summarized texts, do not truly measure the legal capacity —did all required elements of a delivered legal case (handled, step, action) appear in the summed text). In fact, metrics from domain-based astute experts are most needed to evaluate legal summaries.

7. Conclusion

The use of Natural Language Processing (NLP) for the summarization of legal documents has a great potential to change the way legal practitioners, researchers, and even citizens access and understand legal information in India. With the increased number of court judgments, statutes, and legal filing—along with how long and complex these documents

are—it is necessary to include automated solutions that can deliver the core content and factual and legal correctness.

This survey presented both classical and modern summarization approaches. It included extractive methods like Text Rank and BERTSUMEXT, and more recent abstractive models such as BART, T5, and PEGASUS, and even hybrid approaches. The extractive summarization methods outperform in factual correctness, while abstractive and hybrid methods have greater readability and coherence. However, the greater the abstraction, the higher the risk of factual hallucination on the author's behalf. The method selected will depend on the specific use case at hand, whether legal professional, public engagement, or academic research.

References

[1] Vidhi Centre for Legal Policy, Judicial Pendency in India: Challenges and Solutions, 2021.

[2] Indian Kanoon, "Judgments repository."

[3] V. Parikh, V. Mathur, P. Mehta, N. Mittal, and P. Majumder, "Law Sum: A weakly supervised approach for Indian legal document summarization," in Proc. ACM Symp. on Document Engineering (DocEng), 2021, pp. 1–4.

[4] Government of India, "Indian Code—online statutes repository." Avallabile: https://indiacode.nic.in.

[5] Supreme Court of India, "Judgments Portal – Supreme Court of India,"

[6] SCC Online, "SCC Online – Legal Research Platform,"

[7] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proc. Int. Conf. Document Analysis and Recognition* (ICDAR), 2007, pp. 629–633

[8] P. Kapoor and A. Kumar, "Justice delayed is justice denied: Legal AI for bail prediction on Hindi judgments," IIIT Hyderabad Technical Report, 2022.

[9] P. Bhattacharya, A. Paul, and S. Pal, "DELSumm: Extractive summarization of Indian Supreme Court judgments," Journal of Indian Law and Technology, vol. XX, no. X, pp. XX–XX, 2020.

[10] S. Sharma, S. Srivastava, and A. Verma, "Comprehensive analysis of Indian legal document summarization: Extractive and abstractive approaches," SN Computer Science, vol. 4, no. 3, pp. 1–13, 2023.

[11] A. Roy, P. Kapoor, and R. N. Sharma, "HLDC: Hindi Legal Document Corpus for extractive summarization and classification," in Proc. International Conference on Computational Linguistics (COLING), 2023, pp. XX–XX.

[12] G. Bhatia, A. Tewari, and D. N. Giri, "Exploring summarization performance: A comparison of pointer generator, PEGASUS, and GPT-3 models," Nanotechnology Perceptions, vol. XX, no. XX, pp. 2255–2269, 2024, doi:10.62441/nano-ntp.vi.3187.

[13] K. Lee, “Natural language processing for automated legal document summarization,” Journal of Applied NLP, vol. XX, no. XX, pp. XX–XX, 2024.

[14] T. Y. S. S. Santosh, M. Aly, and M. Grabmair, “LexAbSumm: Aspect-based summarization of legal decisions,” arXiv preprint arXiv:2404.00594, 2024.

[15] D. Datta, R. Shukla, and A. Majumder, “MILDSum: Multilingual summarization of Indian legal case judgments,” arXiv preprint arXiv:2310.18600, 2023.

[16] S. K. Nigam, R. Ghosh, and V. Jain, “VidhikDastaavej: Structured legal document generation and summarization in Hindi,” arXiv preprint arXiv:2504.03486, 2025.