LEVERAGING BERT FOR NAMED ENTITY RECOGNITIONIN INDIAN COURT JUDGMENTS

[1]Mr. E. Chitti Babu, [2]Mr. A. Ramesh,
[1,2]Assistant Professor ,Department of Computer Science & Engineering,
Geethanjali Institute Of Science And Technology, Nellore, India

Abstract
As Indian court decisions get more numerous and complicated, it becomes harder for lawyers who need to quickly and accurately analyze legal documents. It takes a lot of work to manually mark up these kinds of documents to find important legal entities, and it can be hard to keep things consistent. This is why automation is so important. To do this, the research focuses on Legal Named Entity Recognition (NER), which is an important step in organizing unstructured legal documents. We fine-tune the RoBERTa language model on a set of Indian court judgments that have been labeled so that it can correctly find domain-specific items including petitioners, respondents, statutes, and judges. The immediate result of our work is an automated system that does a great job of annotating legal papers. our cuts down on the amount of labor that needs to be done by hand and makes it possible to process large amounts of judicial texts. This initiative is also a crucial first step toward creating LegalAI systems that will help with legal search, precedent analysis, and decision intelligence in the Indian judicial system.

INTRODUCTION

Legal practitioners who need quick and accurate information from legal texts have big problems because it is getting harder to process and understand the expanding number and complexity of Indian court decisions. This research tries to solve this problem using an automated Legal Named Entity Recognition (NER) system that uses a finely-tuned RoBERTa language model trained on annotated Indian court judgments to find specific entities in the legal field, like petitioners, respondents, statutes, and judges.

This project is very important since it takes a long time to manually annotate legal documents, and the results are not always the same. Many existing methods don't have enough domain-specific accuracy or automation, which makes them less useful in legal operations. This system cuts down on manual work, makes legal text processing more scalable, and sets the stage for more advanced LegalAI systems in the Indian court system by automating high-quality legal annotation.

In the next parts, we'll look at how this system works, the technology that make it operate, and how its ability to find important legal entities is measured. The goal of this project is to make it easier to understand legal texts, which will help applications like legal search, precedent analysis, and decision intelligence. This will make it easier for everyone to get legal information and provide professionals in the Indian legal field more authority.

MOTIVATION

The reason for this initiative is that it is becoming more and more important to make it easier to analyze Indian court decisions, which are becoming more numerous and complicated. Legal practitioners are finding it harder and harder to get timely and reliable information from large amounts of court text. They typically have to rely on human annotation methods that take a long time and are prone to mistakes. This inefficiency not only slows down the legal process, but it also makes it harder to get important legal information and make smart decisions.

This project intends to give legal professionals and scholars the tools they need to quickly find important parts of court decisions, like petitions, respondents, statutes, and judicial authorities, by

creating an automated Legal Named Entity Recognition (NER) system. Imagine a time when lawyers can quickly get structured, high-quality information from complicated legal papers. This would let them focus on legal reasoning and strategy instead of having to manually pull out data. This initiative is really important for making that future happen.

Using advanced natural language processing methods, including fine-tuning the RoBERTa transformer model on an annotated legal corpus, creates a strong framework for accurately identifying entities in the legal field. This method solves the problems with unstructured legal texts and makes legal document annotation much more accurate, consistent, and scalable. The study shows how these technologies may be used to make the process of finding legal information faster and easier. This sets the stage for future uses including legal search, precedent analysis, and decision intelligence.

This work is really useful for a lot of different people. For example, legal practitioners can use this application to speed up case analysis and get easier access to key legal material. Law companies and courts can use this technology to make their job more efficient and up-to-date. In the end, this project helps the bigger goal of LegalAI, which is to promote new ideas and smart automation in the Indian legal system.

IMPORTANCE OF ACCURATE LEGAL META DATA

Accurate legal metadata is vital for structuring and interpreting the increasingly complex and voluminous corpus of Indian court judgments. In the context of Legal Named Entity Recognition (NER), metadata refers to key legal entities such as petitioners, respondents, statutes, judicial officers, case numbers, and dates. Identifying these entities with precision is fundamentaltotransformingunstructuredlegaltextintostructured,searchable,andanalyzable data.

Reliable metadata significantly enhances information retrieval, allowing legal professionals, researchers,andinstitutionstoefficientlylocaterelevantcaselaw,statutoryreferences,and judicial opinions. It reduces ambiguity in legal texts by ensuring that entities are correctly identified and consistently labeled, minimizing the risk of misinterpretation or oversight.

ThisaccuracyisalsocriticalfortheperformanceofdownstreamLegalAIapplications.Systems such as automated legal search tools, precedent analysis engines, and decision intelligence platformsdependheavilyonclean,consistentmetadatatofunctioneffectively.Accurateentity recognition directly impacts the reliability of such tools, making them more useful for practitioners engaged in high-stakes legal work.

Moreover,accuratemetadatasupportsscalableprocessingofjudicialtexts,makingitpossible to analyze large volumes of legal documents rapidly—something that manual annotation cannot achieve. It facilitates multilingual and multijurisdictional integration, which is particularly important in India, where legal documents may be issued in various regional languages across different courts and legal systems. By standardizing key information, metadata helps unify and cross-reference legal data across languages and jurisdictions.

In essence, metadata extracted through NER forms the backbone of intelligent legal informationsystems.IntheIndianlegaldomain,wherecomplexity,scale,anddiversitypresent unique challenges, accurate legal metadata enables a more efficient, transparent, and insight- driven justice system. It empowers legal professionals to focus on interpretation and strategy rather than data extraction, ultimately contributing to faster, fairer, and more informed legal outcomes.

OBJECTIVE

The primary objective of this project is to develop an advanced system for the automatic extractionandannotationofkeylegalentitiesfromIndiancourtjudgmentsusingstate-of-the-artlanguagemodelssuchasRoBERTa.Thesystemaimstoovercomethelimitationsofmanual legal document analysis, streamlining the identification of critical legal components to help professionals efficiently process complex judicial texts.

Inadditiontoentityrecognition,theprojectintroducesadomain-specificfine-tuningapproach tailored to the

legal context.This method enhances the precision and consistency of extracted entitiesbytrainingonan annotatedcorpusofIndiancourtjudgments,ensuringthattheoutput is not only accurate but also aligned with the specific needs of legal professionals. By doing so, the project significantly reduces manual effort, supports scalable legal document processing, and lays a solid foundation for downstream LegalAI applications such as legal search, precedent analysis, and decision intelligence.

LITERATURE SURVEY

Thischapterreviewskey researchinLegalAI,NamedEntityRecognition(NER),and domain-specificadaptationsoflanguagemodelslikeBERT—allaimedatimprovingmachine understanding of complex legal texts such as court judgments. We highlight influential studies spanning legal NLP, Indian court judgment datasets, legal-domain BERTadaptations, and span-based NER improvements. Together, these works form the foundation for building smarter, more interpretable legal document analysis systems.

How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence "HaoxiZhong,ChaojunXiao,CunchaoTu,TianyangZhang,ZhiyuanLiu,and Maosong Sun"

This paper explores the application of artificial intelligence, particularly natural language processing(NLP),within thelegal domain—commonlyreferred toas LegalAI.Withgrowing interest from both AI researchers and legal professionals, LegalAI has shown promise in reducing the burden of manual legal work. While legal experts typically rely on rule-based, symbolic reasoning, NLP researchers employ data-driven and embedding-based approaches. This work reviews the historical development, current advancements, and future directions of LegalAI. It examines legal tasks through both legal and computational lenses, highlights representative applications, and provides a critical evaluation of existing methods.

ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation "VijitMalik,RishabhSanjay,ShubhamKumarNigam,KripabandhuGhosh,Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi"

Thispaperintroducesthe IndianLegalDocuments Corpus(ILDC),adatasetof35,000Indian Supreme Court cases annotated with court decisions.Asubset includes expert-provided gold- standard explanations. To support the development of automated legal reasoning tools, we definethetaskofCourtJudgmentPredictionandExplanation(CJPE),whichaimstogenerate bothaccurateandinterpretablepredictionsofcaseoutcomes.Weevaluateseveralbaseline models andproposeahierarchical occlusion-basedapproachtoenhanceexplainability.While our best model achieves 78% accuracy—compared to 94% for human experts—the disparity in interpretability between algorithmic and expert explanations highlights the complexity of legal reasoning and the need for further research.

LEGAL-BERT: The Muppets straight out of Law School

"Ilias Chalkid is, Manos Fergadiotis, ProdromosMalakasiotis,NikolaosAletras,andIon Androutsopoulos"

This paper examines the adaptation of BERT models for legal NLP tasks, highlighting that commonly used pre-training and fine-tuning practices may not transfer effectively to specialized domains like law. Through experiments on multiple legal datasets, we evaluate three strategies: (a) using the original BERT model, (b) further pre-training on legal-specific texts, and (c) training BERT from scratch on legal corpora. Our results suggest that careful adaptationisessentialforoptimalperformance.Wealsoadvocateforbroaderhyperparameter tuning during fine-tuning and introduce LEGAL-BERT, a suite of BERT models tailored for legal NLP, aimed at advancing research in computational law and legal technology.

SpanBERT: Improving Pre-training by Representing and Predicting `Spans "MandarJoshi,

DanqiChen,Yinhan Liu, Daniel SWeld, LukeZettlemoyer, and Omer Levy"

This paper introduces SpanBERT, an enhanced pre-training approach aimed at improving span-leveltextrepresentations.UnlikeBERT,SpanBERTmaskscontiguousspansinsteadof individual tokens and trains span boundary representations to reconstruct the full masked span. This method yields significant performance improvements on tasks involving span prediction, such as question answering and coreference resolution. Using the same architecture and training data as BERTlarge, SpanBERT achieves F1 scores of 94.6% on StanfordQuestionAnsweringDataset(SQuAD)1.1and88.7%onSQuAD 2.0.Italsosetsa new benchmark on the OntoNotes coreference task (79.6% F1) and performs strongly on TACRED and GLUE benchmarks.

Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition

"HirokiOuchi,JunSuzuki,SosukeKobayashi,ShoYokoi,TatsukiKuribayashi,Ryuto Konno, and Kentaro Inui"

This study proposes an interpretable method for structured prediction tasks by leveraging instance-based learning. The approach computes similarities between spans and assigns class labels based on similar spans from the training data. This allows users to clearly see the influenceofeachtraininginstanceonagivenprediction.Appliedtonamedentityrecognition, the method achieves strong interpretability while maintaining competitive performance.

PROPOSED SYSTEM

The proposed system focuses on enhancing Named Entity Recognition (NER) for Indian court judgments by developing a baseline model using advanced techniques, particularly leveraging the transformer architecture, specifically BERT-like models along with transition-based parsing. Here is a detailed breakdown of the methods used in the proposed system:

Methods Used in the Proposed System

CORPUS CREATION:

The project introduces a new corpus specifically designed for the legal domain, consisting of 46,545 annotated legal named entities mapped to 14 distinct legal entity types. This corpus is crucial, as it provides the necessary training data tailored to the complexities and nuances of legal texts.

TRANSFORMER-BASEDARCHITECTURE:

A transformer-based model is the main part of the proposed NER system. It has been proved to be better than other models at many NLP tasks. Models like BERT and its variants (RO RoBERTa) are great at figuring out how words are related to each other in context, which is important for comprehending legal language and how entities are related to each other.

TRANSITION-BASEDPARSING:

The enhancement of the transformer architecture with a transition-based parser allows the model to improve accuracy in recognizing entities. This integration helps in better capturingdependenciesandrelationshipswithinthesentencestructure,thus addressing issues inherent in traditional systems.

POST-PROCESSINGTECHNIQUES:

The proposed system includes sophisticated post-processing techniques that enhance the accuracy and context consideration of extracted entities.

This process involves:

EntityReconciliation: Adjustingentityclassificationsbasedonprevious occurrences in the document to prevent misclassification due to context loss.

Coreference Resolution: Linking various references to the same entity through out the text(e.g.,apartymentionedinthepreambleandreferredtolater by a different term).

METHODOLOGY

SYSTEM ARCHITECTURE

The methodology of this project is built upon the idea of enriching legal document understanding through deep Named Entity Recognition (NER), specifically customized for Indian court judgments. The overall architecture is modular and scalable, allowing for streamlinedintegrationbetweenpreprocessing,modelinference,andoutputvisualizationusing anintuitiveinterface.Thearchitecturalpipelineissegmentedintofivelayers,eachresponsible for a crucial part of the end-to-end workflow.

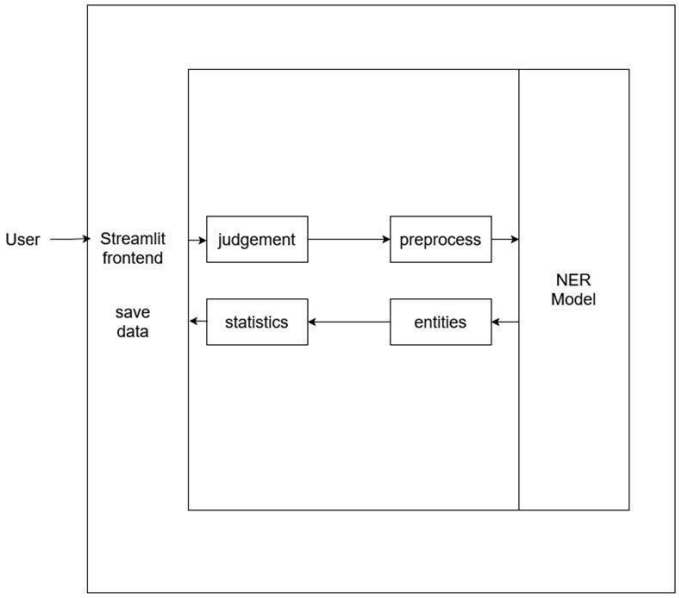

Fig.1. System Architecture

1. **Input Layer:** Accepts text or PDF formats of court judgments.
2. **Preprocessing Layer:** Parses and cleans the text, detects entity-rich sentences, and prepares the data.
3. **NER Processing Layer:** Applies a transformer-based NER model integrated with a transition-based dependency parser for robust entity detection.
4. **Post-processingLayer:**Resolvescoreferencesandassignscontext-awarelabelsusing rule-based strategies.
5. **VisualizationLayer:**RendersresultsinteractivelyviaStreamlit,withcolor-coded entities and expandable summaries.

This structured architecture ensures modularity and supports future enhancement like multi- language support, integration with legal databases, or linking to case laws and judgments via citations.

SYSTEM IMPLEMENTATION

INTRODUCTION

System implementation is the phase where the theoretical constructs of design are translated into a working application. This chapter provides a detailed walkthrough of how the legal Named Entity Recognition (NER) system was developed, the technologies used, how each module was built and tested, and how the components were integrated into a unified pipeline. The implementation follows a layered architecture that includes:

1. TextIngestionandPreprocessing
2. NERModelIntegration
3. Post-processing(CoreferenceResolutionand Reconciliation)
4. InteractiveStreamlit-basedUser Interface
5. PerformanceEvaluationandExport Utilities

Each of these modules is implemented in Python and integrated seamlessly to provide an efficient

and user-friendly experience.

MODULE1: DATA INGESTION AND PREPROCESSING

SUPPORTED FILE TYPES

Thesystemisdesignedtohandlelegaldocumentsintwowidelyusedformats:

PlainText(.txt)

PortableDocumentFormat(.pdf)

This support is a foundational part of the system's architecture, ensuring compatibility with a broad range of legal document sources. The capability to process both formats is seamlessly integrated into theText Ingestion and Preprocessing module — the first layer in the system's pipeline.

PlainText(.txt)

Plaintextfilesprovidealightweight,unformattedstructurethatishighlysuitableforautomated processing. Legal documents in .txt format can be efficiently parsed and tokenized, making them ideal for natural language processing tasks. During the preprocessing stage, these files arequicklyingested,cleaned,andpreparedforentityextraction.Theirsimplicityreduces preprocessingoverheadandenhancesthespeedofdownstreammodulessuchastheNER Model Integration.

PortableDocumentFormat(.pdf)

PDFisthestandardformatforformallegaldocumentationduetoitsabilitytopreservelayout, font styles, and embedded annotations. The system supports both:

NativePDFs(digitallygeneratedandtext-accessible)

ScannedPDFs,whichareimage-basedandrequireOpticalCharacterRecognition (OCR) to extract text ScannedPDFsarepassedthroughOCRtoolsduringpreprocessingtoensurenoinformationis lost,enablingthesystemtoworkwithlegacyorprintedlegaldocuments.Oncetextisextracted, it flows through thesame processing pipelineas .txt files, allowing forentity extraction, post- processing (including coreference resolution and entity reconciliation), and visualization through the Streamlit-based User Interface.

PDFPARSINGIMPLEMENTATION

For PDF parsing, the PyMuPDF (fitz) library was employed due to its high accuracy and its ability to preserve the semantic structure of documents. This library effectively extracts text from multi-page PDF files while maintaining the logical flow and order of content—an essential requirement for downstream tasks such as Named Entity Recognition (NER), where context and sequence play a critical role.

ThesystemprocessesPDFjudgmentsbyopeningthefileandsequentiallyreadingeachpage's content. The extracted text from all pages is concatenated into a continuous block, ensuring that the original narrative and structure of the legal document are retained. This approach is particularlyvaluableinhandlingcomplexlegaldocuments,whereentitiesandreferencesoften span across multiple pages.

By preserving the reading order and document coherence, the extracted raw text serves as a reliable input for the NER model, improving the accuracy and contextual relevance of entity recognition.

TEXTCLEANINGANDPREPARATION

BeforetheextractedtextispassedtotheNamedEntityRecognition(NER)model,itundergoes atextcleaningprocesstoensuretheinputisconsistent,noise-free,andoptimizedforaccurate entity extraction. This preprocessing step involves several key actions:

Removal of Page Headers and Footers:

Repeated elements such as headers and footers, which often appear on each page of legal documents (e.g., court names, case titles, or page numbers), are stripped out to prevent them from interfering with entity recognition.

White space Normalization:

Excessive or irregular spacing is standardized to ensure consistent sentence structure and word boundaries, which helps improve tokenization and model performance.

Elimination of Non-ASCII Characters:

Any characters outside the standard ASCII range are removed to maintain text compatibility and prevent encoding issues. This is especially useful when handling scanned or OCR-processed documents that may introduce unexpected symbols.

By applying these cleaning steps, the system produces a well-structured, uniform text input thatenhancestheperformanceandaccuracyoftheNERmodelinidentifyinglegalentitiesand contextual references.

SENTENCE SEGMENTATION

Once the text is cleaned, it is passed through spaCy, a powerful natural language processing (NLP)library,tosplitthetextintomeaningfulsentencetokens.Thisstepiscrucialforenabling sentence-level tagging, which enhances the granularity and accuracy of entity recognition.

spaCy'slanguagemodel(en_core_web_sm)processesthecleanedtextandidentifiessentence boundaries based on syntactic cues, punctuation, and linguistic structure. This results in a list of well-defined sentences, each treated as an individual unit for analysis.

Bysegmentingthetextintosentences,thesystemsupports:

Context-awaretagging,whereentitiescanbeidentifiedwithrespecttotheir surrounding sentence.

Improvedpost-processing,allowingformoreprecisecoreferenceresolutionandreconciliation across related sentences.

Modularevaluation,enablingthesystemtoanalyze,visualize,orcorrectentitypredictions at the sentence level.

This structured approach to sentence segmentation ultimately improves the quality of NER output and makes downstream tasks more effective.

MODULE2:LEGALNERMODELINTEGRATION

The model used is a RoBERTa +Transition-based Parser pipeline pre-trained on Indian court data. It was loaded using spaCy's transformer-compatible architecture.

LOADINGTHEPRETRAINED MODEL

To perform Named Entity Recognition (NER) on legal texts, the system utilizes a pretrained transformer-based NER model specifically fine-tuned for legal domain language. This model isloadedusingthespaCyNLPframework,whichprovidesastreamlinedinterfaceforapplying advanced NLP models to raw text. The selected model—en_legal_ner_trf—is trained on a corpus of legal documents and is capable of recognizing legal-specific entities such as case names, statutes, court names, dates, and legal roles. By loading this model at runtime, the systemisabletoleverageitslearnedrepresentationsandentityrecognitioncapabilitieswithout requiring additional training.

This pretrained model serves as the core engine for the NER module, enabling the system to tag relevant entities within each segmented sentence and prepare them for further post- processing and user interface

display.

ENTITYPREDICTIONWORKFLOW

Once the text has been cleaned and segmented into sentences, each sentence is individually passedtothepretrainedlegalNERmodelforanalysis.Themodelprocesseseachsentenceand identifies named entities using its trained understanding of legal language. Entity predictions areaccessedthroughthemodel'sdoc.entsattribute,whichcontainsalistofrecognizedentities along with their corresponding labels (e.g., PERSON, DATE, STATUTE, CASE_NAME). These labels are essential for categorizing and interpreting the legal information contained in the document.

The system iterates over all sentences and aggregates the extracted entities into a unified list. ThiscollectionformsthecoreNERoutput,whichcanthenbeusedfordownstreamtaskssuch as coreference resolution, visualization in the user interface, or structured export for further legalanalysis.Byhandlingsentencesindividually,themodelensuresbettercontextualtagging and avoids issues that may arise from processing large blocks of unstructured text all at once.

RESULTS

Fig.2. Home page of Legal NER

Fig.3. Legal NER Input Interface

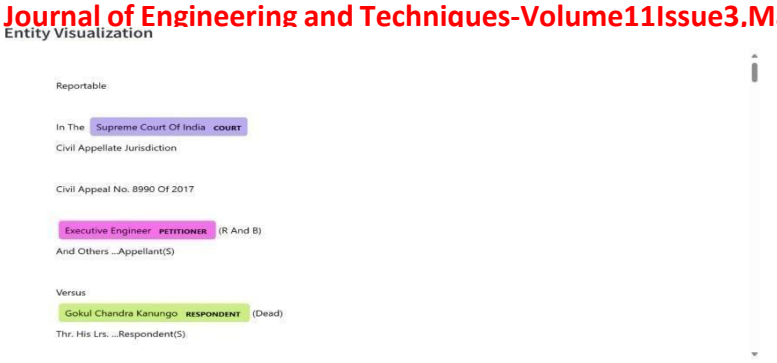

Fig.4. VisualizationofRecognizedEntities-1

Fig.5. VisualizationofRecognizedEntities-2

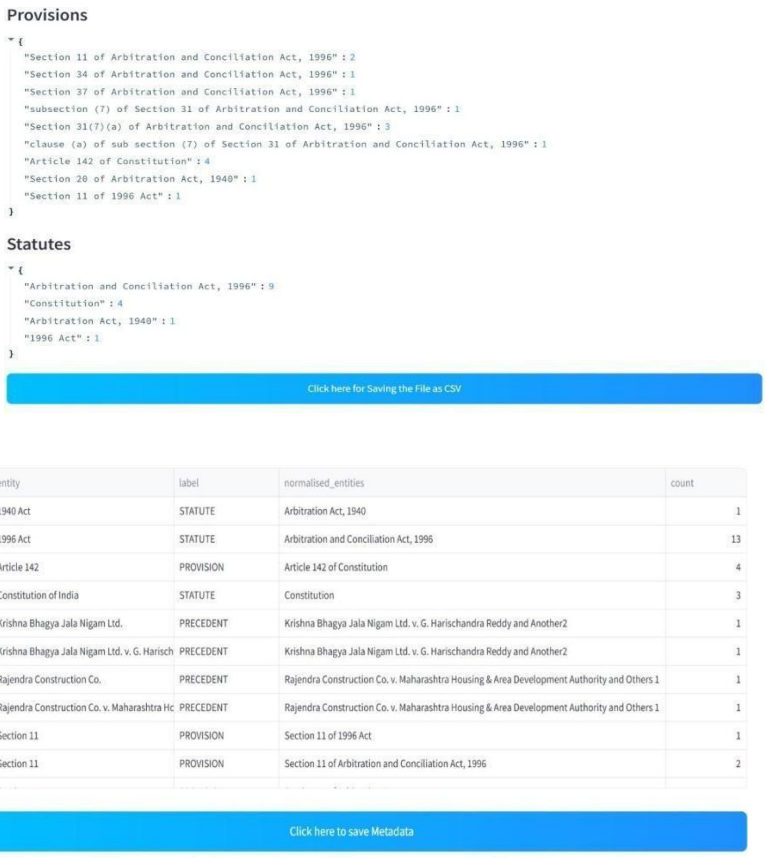

Fig.6.
Save Recognized Entities as CSV file

CONCLUSION

This research shows how to employ a fine-tuned RoBERTa model for Named Entity Recognition (NER) in Indian court rulings, which is an increasing requirement in the legal field for automation. The approach greatly cuts down on the need for manual annotation by reliably identifying important legal entities like petitioners, respondents, legislation, and judicial officers. This makes processing complicated legal texts more consistent, efficient, and scalable. This work not only makes document annotation better right away, but it also sets the groundwork for creating advanced LegalAI applications that are specific to the Indian legal system. It brings up new methods to improve tools for legal research, analyzing precedents, and helping people make decisions. As the amount and complexity of legal data grows, the methods and outcomes of this study show how NLP-driven solutions could change the way legal information systems work.

FUTURE SCOPE

This project establishes a strong foundation for automating legal text analysis in the Indian judicial system using a fine-tuned RoBERTa model for Named Entity Recognition (NER). Futureworkcanfocusonexpandingtherangeofrecognizedentitiestoincludecasenumbers, legal provisions, court names, and case outcomes, enabling a more comprehensive understanding of judgments. Supporting multilingual capabilities and adapting the model to different court jurisdictions across India can further enhance accessibility and relevance. Additionally, integrating the extracted entities into legal knowledge graphs can improve legal research, precedent tracking, and semantic search applications.

The system can also be extended for real-time use in digital courtrooms and legal research platforms,significantlyreducingmanualworkload.IncorporatingexplainableAImethodswill improve transparency and trust, which is crucial in legal contexts. Collaboration with LegalTech platforms can enable seamless integration into existing workflows, while the structured output from NER can support downstream tasks such as judgment summarization, legal question answering, and decision prediction. Finally, implementing continuous learning mechanisms will ensure the model stays current with evolving legal language and practices, driving long-term impact in the development of LegalAI systems.

References

1. HaoxiZhong,ChaojunXiao,CunchaoTu,TianyangZhang,ZhiyuanLiu,andMaosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5218–5230.
2. Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha,Arnab Bhattacharya, andAshutosh 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), pages 4046–4062.
3. IliasChalkidis,ManosFergadiotis,ProdromosMalakasiotis,NikolaosAletras,andIon Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In FindingsoftheAssociationforComputationalLinguistics:EMNLP2020,pages2898– 2904, Online. Association for Computational Linguistics.
4. Mandar Joshi, Danqi Chen,Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy.2020.SpanBERT:ImprovingPre-trainingbyRepresentingandPredictingSpans. Transactions of the Association for Computational Linguistics, 8:64–77.
5. Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, ShoYokoi,Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-Based Learning of Span Representations:A Case Study through Named Entity Recognition. In Proceedings of the 58th Annual Meeting of theAssociation for Computational Linguistics, pages 6452–6459.
6. IliasChalkidis,ManosFergadiotis,ProdromosMalakasiotis,andIonAndroutsopoulos. 2021. Neural contract element extraction revisited: Letters from sesame street. arXiv preprint arXiv:2101.04355.
7. JingLi,AixinSun,JiangleiHan,andChenliangLi.2020.Asurveyondeeplearningfor named entity

recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1):50–70.