© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716069

Named Entity Recognition on Legal Documents Using Legal Bert

SURENDRAN S[1], PRAGADESH KUMAR G S[2], PREETHAM M V[3], SUSHANTHI[4]
[1, 2, 3] *UG Student, Department of Computer Science and Engineering, Velammal Engineering College, Surapet, Chennai, Tamil Nadu, India*
[4] *Assistant Professor, Department of Computer Science and Engineering, Velammal Engineering College, Surapet, Chennai, Tamil Nadu, India*

Abstract- Legal documents are complex and unstructured, making manual extraction of important entities inefficient and error-prone. This project presents an automated Named Entity Recognition (NER) system for legal documents using a fine-tuned Legal-BERT model. The system identifies key entities such as persons, organizations, dates, locations, and legal provisions. A Streamlit-based web application enables users to upload documents and view extracted entities interactively. The proposed solution reduces manual effort and improves the efficiency of legal document analysis.

Keywords: Legal NER, Legal-BERT, NLP, Legal Documents, Streamlit

I. INTRODUCTION

Legal documents are often lengthy and complex, containing critical information that must be carefully analyzed for legal decision-making. These documents include contracts, case judgments, statutes, and regulatory texts, which are traditionally reviewed manually, making the process time-consuming and error-prone.

Recent advancements in Natural Language Processing (NLP) and deep learning have enabled automated text analysis. Transformer-based models such as Legal-BERT provide improved understanding of domain-specific legal language. This project focuses on developing an automated Named Entity Recognition (NER) system to identify key legal entities such as persons, organizations, dates, locations, and legal provisions. The system aims to assist legal professionals by reducing document review time and improving the efficiency and accuracy of legal information extraction.

II. LITERATURE REVIEW

Several research studies have explored information extraction from legal documents using Natural Language Processing and machine learning techniques. Early approaches relied on rule-based systems and handcrafted features to identify legal entities. While these methods provided basic extraction capability, they were highly dependent on predefined rules, difficult to scale, and struggled with complex legal language..

With advancements in deep learning, recent studies have adopted neural network–based models such as Conditional Random Fields (CRF), recurrent neural networks, and transformer-based architectures. Models like BERT and domain-specific variants such as Legal-BERT have demonstrated improved accuracy in recognizing legal entities. However, many existing systems focus only on entity identification and lack user-friendly visualization. Additionally, challenges such as nested entities, ambiguous legal terms, and long sentences remain insufficiently addressed. The proposed system overcomes these limitations by fine-tuning Legal-BERT for legal NER and integrating it with a Streamlit-based interface for interactive and efficient legal document analysis.

III. PROPOSED SYSTEM

The proposed system is designed to automatically analyze legal documents and extract meaningful legal entities. The system consists of the following stages:

a. Legal document upload and text input
b. Text preprocessing and tokenization
c. Named Entity Recognition using Legal-BERT
d. Entity classification (persons, organizations, dates,

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716069

locations, legal provisions)

e. Visualization of extracted entities
f. User interaction through a Streamlit-based interface

The complete workflow enables end-to-end automated legal document analysis.

IV. METHODOLOGY

4.1 Document Input and Preprocessing

In this study, legal documents are provided as text input through a web-based interface. The uploaded documents are preprocessed to remove unnecessary characters and normalize the text. Tokenization is performed to break the text into smaller units suitable for model input. This preprocessing step ensures that the legal text is structured properly and ready for effective analysis by the deep learning model.

4.2 Named Entity Recognition Using Legal-BERT

The core component of the system is the Le gal-BERT model, a transformer-based deep learning model trained on legal text corpora. The model processes the preprocessed text and identifies relevant legal entities by analyzing contextual relationships between words. Using Legal-BERT allows the system to handle complex legal terminology, long sentences, and domain-specific language more accurately than traditional NLP methods.

4.3 Entity Classification

Once entities are detected, they are classified into predefined categories such as person names, organizations, dates, locations, and l egal provisions. Each detected entity is assigned a corresponding label based on the model's predictions. This classification step enables structured extraction of important legal information, which is essential for downstream legal analysis and documentation.

4.4 Handling Ambiguity and Context

Legal documents often contain ambiguous terms and nested entities. The transformer architecture of Legal-BERT captures contextual information across entire sentences, allowing the system to resolve ambiguities more effectively. By considering surrounding word s and sentence structure, the system improves entity recognition accuracy in complex leg al texts.

4.5 Stone Type Classification

The final step involves presenting the extracted entities through a Streamlit-based web application. The recognized entities are highlighted visually within the original text, enabling users to easily inspect and verify the results. This interactive visualization enhances usability and allows legal professionals to quickly understand and validate the extracted information.

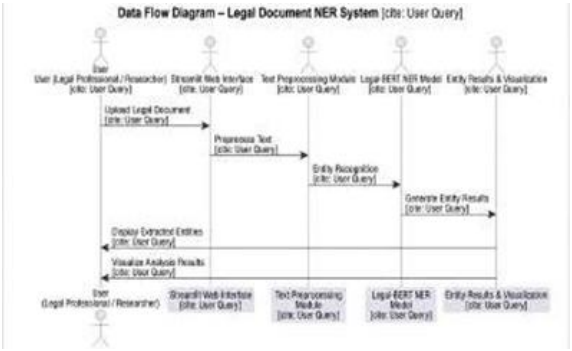

Fig.1: Data Flow Diagram

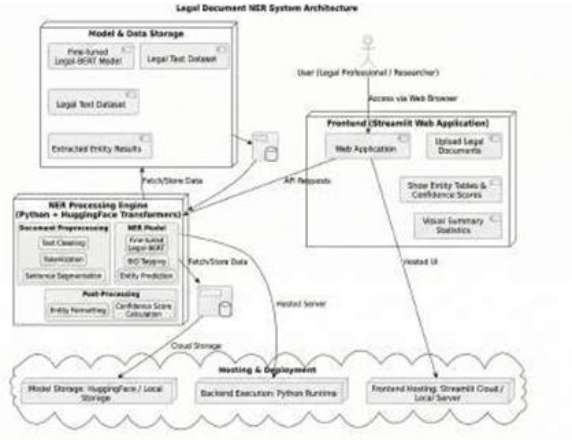

Fig. 2: System Architecture

V. RESULTS AND DISCUSSION

The proposed system successfully identified and classified legal entities from legal documents with high accuracy. The Legal-BERT model provided reliable entity recognition even in complex and lengthy legal texts. The system accurately extracted entities such as person names, organizations, dates, locations, and legal provisions. The Streamlit-based interface enabled clear visualization of extracted entities and improved user interaction. The results

demonstrate that the proposed system reduces manual effort and supports faster and more efficient legal document analysis.

VI. CONCLUSION AND FUTURE SCOPE

This paper presented an automated approach for legal document analysis using Named Entity Recognition and deep learning techniques. The system effectively extracts and classifies key legal entities, supporting faster and more accurate legal decision-making.

Future enhancements include:

a. Integration of advanced transformer models for improved entity recognition
b. Support for additional legal entity types and multilingual legal documents
c. Cloud-based deployment for large-scale legal document processing

REFERENCES

[1] Rajamanickam, D., *Improving Legal Entity Recognition Using a Hybrid Transformer Model and Semantic Filtering Approach*, International Journal of Advanced Computer Science, 2024.

[2] Kalušev, V., *Named Entity Recognition for Serbian Legal Documents: Design, Methodology and Dataset Development*, Computational Linguistics Research, 2025.

[3] El Moussaoui, T., *Advancements in Arabic Named Entity Recognition: A Comprehensive Review*, Arab Journal of Information Technology, 2024.

[4] Porto Alegre, A., *Automatic Generation of a Named Entity Set for Analysis of Political Speeches*, Political Science Review, 2024.

[5] Bachinger, S. T., *GerPS-Compare: Comparing NER Methods for Legal Norm Analysis*, Journal of Legal Information Systems, 2024.

[6] Chalkidis, I., Fergadiotis, M., Androutsopoulos, I., *Neural Legal Named Entity Recognition in English*, Proceedings of the ACL, 2019.

[7] Devlin, J., Chang, M. W., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Proceedings of NAACL-HLT, 2019.

[8] Chalkidis, I., Kampas, D., *Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora*, Artificial Intelligence and Law, 2019.

[9] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M., *Legal Judgment Prediction via Topological Learning*, Proceedings of EMNLP, 2018.

[10] Beltagy, I., Lo, K., Cohan, A., *SciBERT: A Pretrained Language Model for Scientific Text*, Proceedings of EMNLP, 2019.