PROJECT ABSTRACT

Named Entity Recognition for Legal Documents using BERT

Legal-domain NER with a fine-tuned BERT token classifier (BIO tagging): extracts PARTY, DATE, AMOUNT, SECTION and COURT entities from Indian legal text, built on CoNLL-2003 plus ~2,000 annotated legal sentences. Delivered built-to-order with annotation guidelines, notebook, inference API, weights, evaluation report, web demo, PPT and viva kit.

01. Title

Named Entity Recognition for Legal Documents using BERT.

02. Introduction / Background

Lawyers and compliance teams drown in documents where the key facts — who the parties are, the dates, the amounts, the cited sections, the jurisdiction — are buried in prose. Named Entity Recognition automates the first pass: finding and typing these spans. General NER (CoNLL-2003: person/organization/location) misses legal specifics like ‘Section 11’ or ‘Rs. 45,00,000’ in Indian grouping. This project fine-tunes BERT for token classification with the BIO scheme on a legal schema, starting from CoNLL-2003 and adding ~2,000 annotated Indian legal sentences.

03. Problem Statement

Contract review, due diligence and case-law research all begin with entity extraction, done today by junior lawyers reading line by line. Off-the-shelf NER models trained on news text mislabel legal entities and choke on Indian formats (lakh/crore grouping, Act citations). This project builds a legal-specialized extractor: five entity types that matter in Indian agreements and notices, with entity-level precision/recall/F1 measured on a held-out legal test set and an annotation-highlighting demo.

04. Objectives

- Define the legal schema: PARTY, DATE, AMOUNT, SECTION, COURT with annotation guidelines.
- Annotate ~2,000 Indian legal sentences (agreements, notices) in BIO format.
- Fine-tune bert-base-uncased for token classification (3–5 epochs, lr 3e-5).
- Evaluate entity-level precision/recall/F1 on the held-out legal test set.
- Deliver an inference API + web demo highlighting entities in uploaded legal text.

05. Proposed Methodology

(1) Annotation: two-pass labeling with the shipped guidelines; disagreements adjudicated, inter-annotator agreement reported. (2) The base model is bert-base-uncased with a dropout + linear token-classification head over the BIO label set. (3) Training: fine-tune on the legal set (optionally after CoNLL-2003 warm-up), 3–5 epochs, batch 16, lr 3e-5, with validation F1 for early stopping. (4) Decoding: WordPiece sub-tokens are merged to word-level tags; adjacent B-/I- spans become entities. (5) Evaluation uses strict entity-level F1 (exact span + type match).

06. System Architecture / Working

Legal text is WordPiece-tokenized and passed through BERT; each token's contextual embedding is classified into BIO labels. The demo merges sub-word tags, renders the document with color-coded entity highlights and a sortable entity table with per-entity softmax confidence. Indian-specific handling — lakh/crore amount patterns, ‘Section N’ and ‘Act, Year’ citation shapes — is documented as rule-assisted post-processing over the neural tags, an honest hybrid the report defends.

07. Hardware / Software Requirements

- Python 3.10+, PyTorch, Hugging Face transformers
- CoNLL-2003 (base) + annotated legal set (ships with project)
- GPU recommended for fine-tuning; inference runs on CPU
- Flask/FastAPI inference service; HTML/CSS/JS demo

08. Expected Outcomes

A legal-domain NER model, entity-level F1 measured on the held-out legal test set, the annotation guidelines and labeled set, and a highlighting demo. The report breaks F1 down per entity type — SECTION and AMOUNT typically trail PARTY — giving the student a real error analysis to present.

09. Applications

- Contract abstraction and due-diligence triage
- Case-law research (party/court extraction)
- Compliance document processing

10. Future Scope

- Relation extraction (party→obligation links)
- More entity types: CLAUSE, PENALTY, GOVERNING_LAW
- Multilingual legal text (Hindi/regional)