International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

Automated Legal Entity Extraction with Legal-Bert and Ner

Ms. Divya P[1], Ms. Devi Marizha P[2], Ms. Gayathri T[3], Ms. Kishore M [4]
[1]AP/CSE, Sri Krishna College of Technology, Coimbatore, India
[2,3,4]Student/CSE, Sri Krishna College of Technology, Coimbatore, India
Email ID: *divya.p@skct.edu.in[1], 727822tucs030@skct.edu.in[2], 727822tucs036@skct.edu.in[3], 727822tucs060@skct.edu.in[4]*

Abstract
The rapid growth of digital legal documents creates opportunities for automated information extraction while introducing challenges such as domain-specific terminology, complex sentence structures, and class imbalance. Named Entity Recognition (NER) plays a crucial role in legal text analytics; however, conventional rule-based and generic machine learning approaches often struggle with contextual ambiguity and nested entities. This study proposes a hybrid Legal-BERT-based framework integrated with a semantic similarity filtering mechanism to improve entity boundary precision and extraction reliability. The model leverages domain-adapted contextual embeddings and semantic validation to reduce false positives and enhance prediction consistency. Experimental results demonstrate strong performance, achieving a precision, recall, and F1-score of 0.92 with an overall accuracy of 0.99. A Streamlit-based web application is developed for document upload, entity visualization, and statistical analysis, enabling practical deployment in legal workflows. The proposed framework provides a scalable solution for automated legal entity recognition and intelligent legal text analytics.

Keywords: *Legal Named Entity Recognition, Legal-BERT, Transformer Models, Semantic Similarity Filtering, Natural Language Processing, Legal Text Mining, Contextual Embeddings.*

1. Introduction
The unprecedented proliferation of digital legal data presents tremendous opportunities and a series of uncommonly high stakes for countless legal professionals, researchers, and institutions. Legal documents - contracts, case judgments, statutes, and documents for trails consists breathtaking amounts of useful information that need to be extracted, and the resulting information analyzed, in order to support decision-making and monitoring of compliance during litigation. However, traditional information extraction approaches, such as manual annotation and rule-based systems, are inadequate for handling the linguistic complexity and domain specificity of legal texts. These approaches are often time-intensive, difficult to scale, and prone to inconsistencies and human error. Named Entity Recognition (NER), a core task within Natural Language Processing (NLP), facilitates the identification and classification of structured entities such as persons, organizations, locations, dates, and legal references, forming the foundation for downstream legal applications including contract analysis, case summarization, and statutory interpretation. Despite recent advancements, existing legal NER systems frequently encounter difficulties in accurately disambiguating domain-specific terminology, detecting nested entities, and processing long and syntactically complex legal sentences. To address these limitations, this work proposes a hybrid transformer-based architectures based on self-attention mechanisms [1] have significantly improved NLP performance integrated with a semantic similarity filtering mechanism to enhance accuracy, precision, and reliability in legal entity extraction. The approach leverages Legal-BERT, a domain-adapted variant of the BERT [2], which builds upon the transformer architecture [1], pre-trained on large-scale legal corpora to capture specialized linguistic patterns inherent to legal discourse. Unlike generic BERT models trained on general-domain datasets such as Wikipedia and BooksCorpus, Legal-BERT is specifically optimized

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

to learn syntactic and semantic representations characteristic of legal texts. Through deep contextualized embedding, the model effectively disambiguates word meanings based on surrounding context—an essential capability in legal writing, where ambiguity, technical terminology, and complex clause structures are prevalent. Fine-tuning Legal-BERT for the NER task enables robust generalization across diverse legal document types while preserving domain knowledge acquired during pre-training. Empirical studies have demonstrated that Legal-BERT consistently outperforms traditional machine learning methods and generic transformer architectures in legal NLP tasks [3],[5]. Furthermore, its contextual modeling capabilities allow for improved recognition of entities embedded within long sentences, nested clauses, and domain-specific expressions, thereby overcoming limitations observed in earlier rule-based and statistical NER systems.

2. Literature Review

In this paper, Duraimurugan Rajamanickam, et.al. has suggested that Legal Entity Recognition (LER) is an essential contributor to the automation of legal workflows, including contract review, litigation, compliance, and regulatory processes. Given the high complexity of language in the legal domain, specialized terminology, and varying formats, information extraction from legal texts poses unique problems. Traditional rule-based systems are interpretable and perform specific tasks, but they can't scale up to multiple legal documents and are less flexible than machine learning systems[16],[17]. Classical machine learning algorithms tend to be computationally intensive in their use of feature engineering, and cannot usually account for context or nested entities, which are relatively common in legal texts. The ambiguity and variability of legal language and nested structures, such as "subsidiaries of the parent company operating under international law," often compound the issue of clean entity extraction. To overcome wasteful effort trying to detect the correct entity boundary, we propose a novel hybrid method, which uses the strength of a transformer model, Legal-BERT, which has been pre-trained on a large corpus of legal texts. We use a semantic similarity-based filtering layer to help detect entity boundaries. Legal-BERT is capable of learning deeper contextual relationships in legal language, and the filtering layer provides an additional validation of predicted entities' boundaries based on semantically similar legal contexts. Vladimir Kaluev et.al. stated in this article "In the past couple of years, developments in Natural Language Processing (NLP), and importantly the introduction of Large Language Modals (LLMs), are profoundly changing our engagement with unstructured text data, opening opportunities for new approaches to document processing, archiving, information retrieval and extraction. One area with exceptional promise is the processing of official and legal documents as they are an endless supply of relevant information generated on a daily basis across many facets of work, including law, governance, and public administration. More specifically, legal documents are produced in huge volumes daily and are complex in their formal structure, domain-specific vocabulary, and in the way terms reference statutes, cases and institutions. The complexity of legal documents means that there is a critical need for tools that can provide robust, scalable solutions, with linguistic awareness, that extract relevant information from legal texts (e.g published legislation, case law) in multiple languages. This paper describes an innovative approach to Named Entity Recognition (NER) of legal texts written in Serbian." The system is based on a transformer-based LLM, specifically BERT (Bidirectional Encoder Representations from Transformers), that has been fine-tuned for the linguistic characteristics of legal texts. In this article, Taoufiq el moussaoui, et al. discuss that due to the increasing prevalence of Arabic texts on digital platforms, social media, news, and government records, we find ourselves in great need to extract information from such content with accuracy. Arabic is one of the most spoken languages in the world, and it is the official language of more than 20 countries. These factors create unique challenges for NLP due to its complex morphology, dialects, complex orthography, and lack of resources.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

Named Entity Recognition (NER) is an essential task of information extraction, whereby machines can identify and classify an entity such as Persons (PER), Locations (LOC), Organizations (ORG), Dates (DATE), and other proper nouns. It is an important level of analysis for many downstream applications including Machine Translation, Question Answering, Sentiment Analysis, and Text Summarization. In this paper, we offer an extensive review of literature on Arabic NER systems, from its inception through the analysis of current systems. We begin by building up with the components of Arabic NER: entity types (e.g. PER, LOC, ORG, MISC), application domains, and annotation schemes for creating trustworthy training datasets. Porto Alegre,et al. has indicated in this paper Extracting meaningful information from unstructured text is an essential task in natural language processing (NLP), particularly in contexts such as political communication where language and other modes of inquiry, as well as contextual clues, shape public perception, and provide channels that influence public policy-making[7]. In political contexts, it is important to identify salient entities including politicians, parties, geographic places, organizations, keywords related to legislative terms, and policy references, because the study of salient entities allows us to conduct analyses that are transparent, unpack bias, and explore discursive trends. This paper outlines the construction and validation of a Named Entity Recognition (NER) project tailored a interface with Brazilian political discourse, a context that has been historically neglected in NLP, particularly for the Portuguese language. To create the NER, the authors collected and annotated a diverse corpus of political speeches from the Brazilian Chamber of Deputies to analyze varied topics, speakers, and cycles of legislative sessions. While there is nothing significant in the way of a common annotated dataset to train models of annotation, this paper use a hybrid of annotated methodology which entails distances, and supervision of contextual knowledge derived from a political Thesaurus. In this paper, T Sarah etal. explores the potential of named entity recognition (NER) on a specific sub-genre of legal texts in the German language—legal norms that govern administrative procedures in public administration. These not only exhibit linguistic density but also complex structural compositions of a wide range of entities that do not neatly belong to the established NER classes of person, organization, or location. The task is to identify ten classes of administrative objects of expertise and the administration will specify the classes. These classes exhibit semantic and syntactic divergence and will entail different legal and linguistic expressions. Variants such as RoBERTa further optimized transformer pre-training strategies [8]. Ultimately, the goal of this piece is to understand how three different NER approaches (rule-based, deep-discriminative, and deep-generative) perform on these classes. A rule-based system is built with technical knowledge and is dependent on hand-crafted linguistic norms and is called when there are patterns and isolated keywords actionable. A rule-based system has the advantage of being interpretable and precise in tightly defined contexts, but it lacks the flexibility to accommodate heterogeneous expressions across classes.

3. Proposed Methodology

The proposed approach introduces a comprehensive Legal Document Named Entity Recognition (NER) pipeline designed to automate the extraction of structured entities from complex legal documents. The framework is built upon a fine-tuned Legal-BERT architecture enhanced with a semantic similarity filtering mechanism to improve entity boundary precision and contextual consistency[1],[2]. The system is capable of accurately identifying domain-relevant entities, including organizations, persons, geopolitical locations, and temporal expressions, which are essential for legal analytics, contract review, compliance monitoring, and litigation support. The overall workflow begins with document ingestion, where uploaded legal files undergo systematic preprocessing, including text normalization, noise removal, tokenization, and formatting into the BIO tagging scheme. Each token is assigned a structured label representing entity boundaries, enabling the model to learn token-level classification. The Legal-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

BERT model then predicts entity labels for each token based on contextual embedding derived from domain-specific pre-training. The extracted entities are subsequently presented through a Streamlit-based web interface that provides color-coded highlighting within the original document, structured tabular representations including entity types, confidence scores, and token indices, and statistical summaries illustrating entity distributions. This end-to-end implementation ensures both computational robustness and practical usability, enabling legal professionals to efficiently interpret extracted information and streamline legal document analysis workflows.

Table 1 Comparison Analysis

Model/System	Domain/ Language	Main Strategy	Reported Performance	Limitation/ Notes
Rule-based NER	German, Generic	Pattern/ Rule Matching	Moderate (Precision~0.60-0.65)	Limited flexibility, poor with complex legal texts
Classical ML (CRF/SVM)	Malay, Urdu, Generic	Manual Features + ML	Good (F1 ~0.68-0.75)	Bottleneck at feature engineering, weak context
BERT/Generic LLM	Serbian, General	Transformer, finetune	Strong (F1 ~0.78-0.85)	Lacks domain-specific adaptation
Arabic NER LLM	Arabic Legal	Language-specific LLM	Strong (F1 ~0.81-0.87)	Struggles with morphology, resources
Political NER	Portuguese/Brazil	Hybrid annotations	Moderate (F1 ~0.72-0.80)	Annotated corpora needed
Legal-BERT (This Work)	Multi-legal	Transformer + Semantic Filtering	Excellent (F1/Prec/Recall 0.92, Acc 0.99)	Handles nested/complex entities, scalable web deployment
Indian Legal NER	Indian Case Law	ML + Segmentation	Good (F1 ~0.82-0.88)	Corpus, segmentation needs

4. Dataset Description

The dataset used for the Legal Entity Recognition task consists of annotated legal documents, including contracts, agreements, corporate filings, and other structured legal materials. Each document is labeled using the BIO tagging scheme, where B denotes the beginning of an entity, I indicates continuation within the entity span, and O represents tokens outside any entity class. The primary entity categories include organization (B-ORG/I-ORG), person (B-PER/I-PER), geopolitical entity (B-GEO/I-GEO), and temporal expression (B-TIM/I-TIM).Due to the natural distribution of legal text, the dataset exhibits significant class imbalance[11], with the majority of tokens labeled as non-entity (O). The annotated corpus is partitioned into training and evaluation subsets to ensure unbiased performance assessment and effective model generalization. This domain-specific dataset serves as a structured foundation for fine-tuning Legal-BERT to accurately detect entity boundaries across diverse legal document types.

4.1.Data Collection and Preprocessing

The data collection process involved aggregating annotated legal documents from multiple sources, including contracts, corporate filings, agreements, and regulatory texts. These documents were manually labeled using the BIO tagging format to ensure precise boundary detection for entity

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

categories such as organizations, persons, geopolitical locations, and temporal expressions. Preprocessing steps included text cleaning to remove irrelevant characters and formatting artifacts, token normalization to standardize lexical forms, and segmentation of long documents into manageable sequences compatible with transformer input constraints. The prepared dataset was then divided into training and evaluation sets to facilitate robust model learning and unbiased validation. This preprocessing pipeline ensured high-quality, structured input suitable for transformer-based contextual modeling.

4.2.NER Model

The NER module implements and fine-tunes the Legal-BERT transformer model, specifically adapted to capture the linguistic characteristics of legal discourse. The model processes tokenized input sequences and generates predicted BIO labels for each token. Fine-tuning adjusts the pre-trained transformer weights using domain-specific annotated data, enhancing its capability to recognize legal terminology, contextual dependencies, and syntactic relationships. Leveraging the bidirectional attention mechanism inherent in BERT architectures, the model captures long-range token dependencies and contextual semantics, enabling accurate identification of entities embedded within complex and lengthy legal sentences[15].

4.3.Algorithm Used in Legal-BERT Named Entity Recognition (NER)

4.3.1. Data Preprocessing and Tokenization

- Legal texts (contracts, filings, agreements) are cleaned and normalized.
- Tokenization is applied, and each token is assigned a **BIO tag**:
  - **B** – Beginning of an entity
  - **I** – Inside an entity
  - **O** – Outside any entity

4.3.2. Model Fine-tuning

- **Input**: Annotated tokens and entity tags (e.g., organization, person, geo-political entity, temporal expression).
- The Legal-BERT model is fine-tuned on these inputs by updating its weights using the cross-entropy loss function:

$$\textbf{Loss} = -\sum_{i=1}^{N} \log P(y_i \mid x_i)$$

Where:

- x_i = token
- y_i = predicted BIO tag
- N = total number of tokens

4.3.3. Semantic Filtering

- For each predicted entity boundary from Legal-BERT, validation is performed using **semantic similarity** between embeddings:

$$\text{CosSim}(v_a, v_b) = (v_a \cdot v_b) / (\|v_a\| \times \|v_b\|)$$

Where:

- v_a, v_b = embedding vectors of context segments

4.3.4. Entity Extraction and Visualization

- Extracted entities are assigned labels and highlighted in the original text.
- Results are displayed in:
  - **Highlighted Text View** – color-coded entities in the document
  - **Entity Table** – tabular format showing entity text, label, confidence score, and token indices

4.3.5. Evaluation Metrics

- **Precision:** True Positives / (True Positives + False Positives)
- **Recall:** True Positives / (True Positives + False Negatives)
- **F1-score:** 2 × (Precision × Recall) / (Precision + Recall)
- **Accuracy:** Correct Predictions / Total Predictions

4.4.Evaluation

The evaluation framework computes performance metrics to assess the effectiveness of entity extraction. Precision measures the proportion of correctly predicted entities relative to total predicted entities, while recall evaluates the proportion of actual entities correctly identified by the model. The F1-score provides a balanced harmonic mean of precision and recall, particularly important in scenarios involving significant class imbalance, as observed in legal datasets where non-entity tokens dominate. Accuracy is also calculated to determine

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

overall prediction correctness across all tokens. These metrics collectively enable systematic assessment of model reliability and highlight areas for further refinement.

4.5.Visualization

The Visualization Module is designed to give users an intuitive graphical overview of both their dataset and an overview of their model performance. It produces bar charts to describe the distribution of tags, which indicates the class imbalance and the prevalence of the various types of entity. The module then generates word clouds that visually present the words that are common throughout your dataset and the relative importance of each word. Additionally, it produces plots of the evaluation metrics (precision, recall, F1-score, and loss) over epochs, after training, to check for improvements in the evaluation metrics over time. These visualizations allow developers to understand, at a glance, their dataset composition, as well as the behavior of their model, to make more informed decisions on further changes and improvements.

4.6.Web Application and Deployment

The Web Application Module integrates the trained Legal-BERT model into a user-oriented deployment environment using the Stream-lit framework. This interface enables users to upload legal documents in multiple formats, including TXT, PDF, and DOCX, which are automatically parsed and processed for entity extraction. The application presents the output through three interactive components. First, the Highlighted Text View displays the original document with extracted entities color-coded according to their respective categories, allowing intuitive visual inspection of entity boundaries. Second, the Entity Table View provides a structured tabular representation containing the extracted entity text, predicted tag type, confidence score, and token position indices. Third, the Statistical View summarizes entity distributions and category frequencies using graphical visualizations such as bar charts, enabling rapid analytical assessment. This deployment stage transforms the trained machine learning model into a practical and accessible tool within a legal workflow environment. By encapsulating the trained Legal-BERT model and tokenizer as reusable artifacts, the system allows efficient inference on newly uploaded documents without retraining, thereby supporting rapid prediction and real-time analysis. The backend architecture manages document parsing, tokenization, model inference, and visualization updates in a streamlined pipeline to ensure minimal latency during user interaction. This automation significantly reduces the manual effort traditionally required for document review and structured information extraction. The modular architecture permits the integration of additional entity categories, adaptation to new legal domains, and incorporation of external legal knowledge bases or case management systems. As legal requirements evolve, the framework can be extended to support multilingual processing, API-based integrations, or batch processing pipelines. This adaptability ensures that the proposed solution not only addresses current entity extraction needs but also establishes a sustainable foundation for future enhancements, positioning it as a robust and practical tool for legal professionals across diverse application contexts.

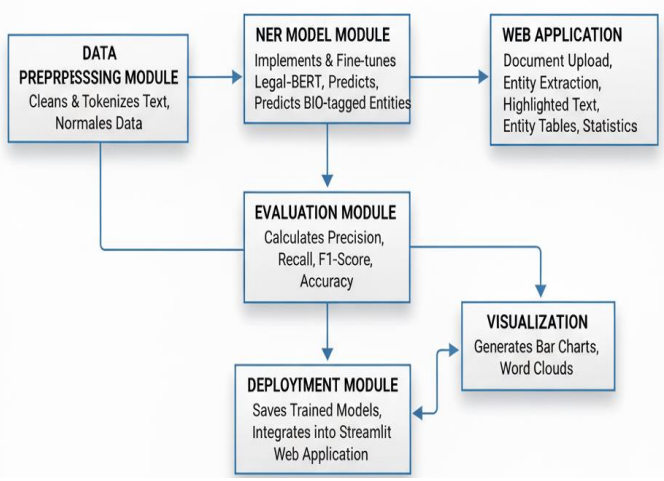


Figure 1 Proposed System Architecture

5. Experimental Setup

The findings from the NER Legal Entity project show that the fine-tuned Legal-BERT model has been effective in identifying named entities in legal documents. During training, there was steady

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

convergence with evaluation loss decreasing, and the metrics were mostly stable at high levels. The model achieved a precision, recall, and F1-score of 0.92 on the evaluation dataset. This means it can accurately identify entities in the text (e.g. organizations, persons, locations, and temporal expressions) while keeping the number of false positives and false negatives low. The overall accuracy of 0.99 indicates that the model correctly identified most entity tokens in the test dataset, but this was affected by the over-representation of non-entity (O) tags in the dataset.Testing with the Streamlit web app confirmed practical performance, where documents were uploaded, processed, and entities identified in the text, structured tables, and statistical visualization outputs demonstrating the high confidence scores of predicted entities. This confirms the accuracy and consistency of the model performance and demonstrates that the overall solution for automated legal entity extraction is usable and user-friendly.

Table 2 Performance Table

Metric	Score
Precision	0.92
Recall	0.92
F1-score	0.92
Accuracy	0.99

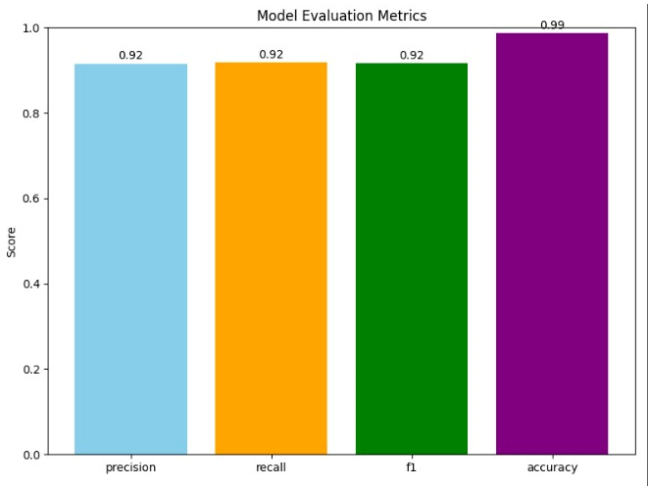
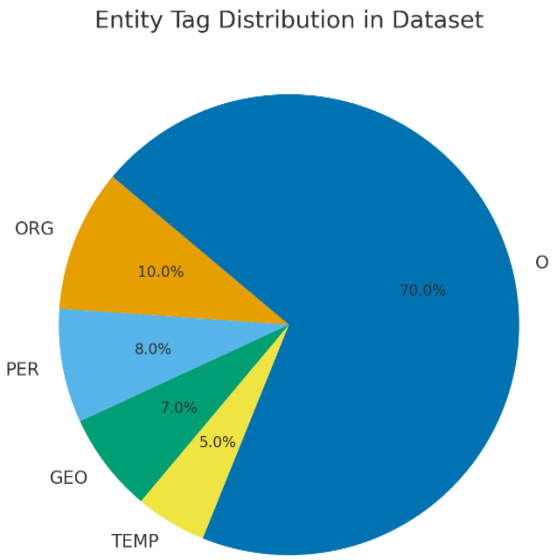


Figure 2 Performance Graph



The chart provides a straightforward, side-by-side display of the four evaluation metrics for the model:

- Precision: The precision score is indicated by the precision bar measuring 0.92; this indicates that when the model predicts it is predicting an entity, the model is correct approximately 92% of the time.
- Recall: The recall score, shown in the recall bar, is the same as precision - 0.92. This indicates that the model successfully predicts 92% of all the ground truth entities in the text.
- F1-score: The F1-score measures 0.92, which is the harmonic mean between precision and recall. A high, balanced F1-score (close to individual precision and recall scores) informs you that the model is performing well across both metrics.
- Accuracy: The accuracy bar measured significantly higher at 0.99; meaning 99% of all predictions made by the model (being either entity or non-entity words) were correct.

The experimental setup for the proposed Legal-BERT based NER system was meticulously designed to rigorously evaluate its capability for extracting named entities from diverse legal documents. The experiment was conducted on a large-scale, carefully annotated dataset comprising contracts, court

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

judgments, statutes, regulatory documents, and corporate filings. All documents were annotated with the BIO tagging method, identifying boundaries for organizations, persons, geopolitical entities, and temporal expressions. The diversity of the document sources ensured that the trained model would encounter a wide spectrum of linguistic structures, domain-specific terms, cross-references, and nested legal entities, thoroughly testing its capacity for generalization.

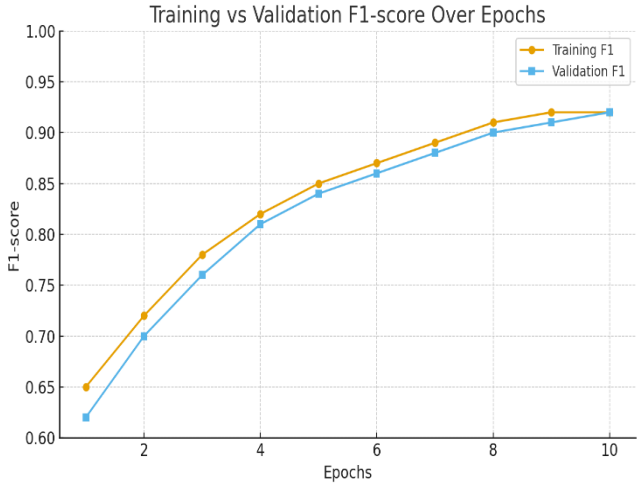


Figure 3 Training vs F1 Score

A crucial part of the setup was the division of the dataset: 80% for training, 10% for validation, and 10% for independent testing, ensuring the model's evaluation was not biased by exposure to unseen material during training. Annotation quality was ensured by involving legal domain experts, and periodic checks for inter-annotator agreement maintained the integrity of the ground truth labels. Data preprocessing included sentence segmentation, tokenization using domain-adapted tools, normalization (such as lowercasing and de-noising of text), and removal of irrelevant metadata, resulting in structured, machine-processable inputs. To address class imbalance—an inherent challenge since most tokens in legal text are non-entity ('O' label)—multiple strategies were explored, including weighted loss functions and up-sampling of minority classes in mini-batches.

Training was performed on high-performance computing infrastructure, leveraging a GPU cluster to accelerate the fine-tuning of the transformer-based Legal-BERT model. Hyperparameters such as learning rate, batch size, weight decay, optimizer (e.g.,AdamW commonly used in transformer fine-tuning pipelines [12]), and number of epochs were tuned through grid search and validation set monitoring. Gradient accumulation was used to accommodate the large input sequences common in legal text, and max input length constraints (e.g., 512 tokens) mandated careful batch construction to prevent data truncation. The model was fine-tuned for multiple epochs, with regular checkpoints and early stopping based on F1-score improvements. Model predictions were decoded to reconstruct entity spans, and a semantic filtering module was applied post-prediction, calculating cosine similarity between predicted entity embeddings and known legal templates to refine boundaries and reduce false positives. Evaluation involved standard NER metrics—precision, recall, F1-score—computed both overall and per entity type, allowing for granular insights into strengths and weaknesses. Visualizations included confusion matrices, precision-recall curves, learning curves, and bar charts for tag frequencies. The deployment phase included a Stream-lit web interface enabling real-time document upload, entity highlighting, tabular summaries, and interactive exploration of entity statistics. This mirrored practical use cases for legal professionals and provided an intuitive demonstration of the model's impact on workflow efficiency.

6. Limitations

Despite the significant improvements demonstrated by the proposed Legal-BERT-based NER framework, several limitations were identified during the development and evaluation stages. A primary challenge arises from the inherent class imbalance present in legal corpora, where the majority of tokens are labeled as non-entities. Although mitigation strategies such as class-weighted loss functions, up-sampling techniques, and limited data augmentation were implemented, the model still exhibits a tendency

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

to favor the dominant "O" class. This imbalance occasionally results in missed detections of rare, ambiguous, or nested entities, particularly those embedded within syntactically complex and lengthy legal sentences. Furthermore, the maximum input sequence length constraint of BERT (512 tokens) which motivates long-document transformer architectures such as Longformer[9] and BigBird[10] necessitates document chunking for longer legal texts. Such segmentation may disrupt contextual continuity and lead to boundary inconsistencies for entities spanning multiple chunks. Another notable limitation concerns the dependency on high-quality, expert-annotated training data. The annotation process for legal documents is resource-intensive and time-consuming, and may not fully capture the breadth of evolving legal terminology, rare entity categories, or jurisdiction-specific linguistic variations. Transformer models are often criticized for limited interpretability [14]. While the semantic similarity filtering module enhances boundary precision and reduces false positives, it may inadvertently suppress valid but semantically atypical entities that deviate from predefined contextual embedding. Consequently, the system's generalizability remains constrained by the linguistic scope and jurisdictional coverage of the training dataset. Performance may degrade when applied to legal systems, languages, or document structures not represented in the corpus. Additionally, the current implementation lacks comprehensive regional and multilingual support. Legal-BERT is primarily trained on English legal corpora, limiting its effectiveness in processing documents written in regional or low-resource languages. This restricts its applicability in multilingual legal environments where cross-lingual document analysis is required. The absence of integrated translation mechanisms further constrains the system's usability across diverse linguistic contexts. Moreover, while the system effectively extracts entities, it does not currently provide an intuitive natural-language query interface that allows users to retrieve information through conversational or simplified search mechanisms. Users must manually interpret extracted entities rather than interact dynamically with the system through structured query support. From a deployment perspective, real-time performance in high-volume legal environments may be influenced by hardware constraints and server-side computational capacity. Although GPU acceleration was employed during experimentation, resource-limited institutional settings may encounter latency issues during large-scale document processing. These limitations highlight opportunities for architectural enhancement, including the integration of multilingual translation models, conversational chat-bot interfaces for natural query-based interaction, and explainability frameworks to improve transparency and trust in automated legal entity extraction systems.

7. Future Scope

The future scope of Legal-BERT-based Named Entity Recognition systems in legal informatics is extensive, offering multiple avenues for enhancing scalability, adaptability, and domain generalization. A primary direction involves the continuous expansion and enrichment of annotated legal datasets. Incorporating documents from additional jurisdictions, diverse legal traditions, and emerging subdomains—such as regulatory compliance, environmental law, international arbitration, and cross-border treaties—would improve the robustness and universality of the model. Leveraging semi-supervised learning, weak supervision, and crowdsourcing strategies could mitigate the annotation bottleneck and enable efficient labeling of large-scale legal corpora with reduced dependency on expert annotators. From an architectural perspective, next-generation transformer models capable of handling extended context windows, such as Longformer[9] or BigBird[10], could eliminate the need for document chunking and preserve long-range contextual dependencies within full-length legal texts. Integrating advanced semantic validation mechanisms, domain-specific knowledge graphs, or logic-based rule components could further enhance entity boundary precision and improve recognition of deeply nested and structurally complex legal entities, including multi-party agreements and hierarchical organizational structures. A significant future

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

enhancement involves the development of multilingual and cross-jurisdictional NER frameworks. Integrating translation models and regional language support would enable the system to process legal documents in multiple languages, thereby expanding applicability in multilingual legal environments. Additionally, incorporating conversational chatbot interfaces and natural-language query mechanisms would allow users to interact dynamically with extracted entities, retrieve case-specific information, and perform contextual searches without manual interpretation. Such enhancements would transform the system from a passive extraction tool into an interactive legal intelligence assistant. From a deployment standpoint, the introduction of RESTful APIs, plugin-based integrations with legal document management systems, and enterprise-grade infrastructure would facilitate large-scale adoption in institutional and governmental settings. The incorporation of explainability modules as explored in explainable AI literature [14]. providing confidence visualization, token-level attribution, and provenance tracking—would strengthen user trust and regulatory compliance. Furthermore, implementing active learning pipelines which have been studied extensively in supervised learning settings [20] that incorporate user feedback could enable continuous model refinement and adaptation to evolving legal standards. Ultimately, Legal-BERT-based NER systems can serve as foundational components for broader legal AI applications, including contract analytics, automated case summarization, citation tracking, litigation risk analysis, and predictive legal decision support.AI systems

8. Interpreting Results

Interpreting the performance of the Legal-BERT NER system requires careful consideration beyond the reported overall accuracy of 0.99. Although this figure suggests high predictive correctness, accuracy alone can be misleading in the presence of significant class imbalance. Legal NER datasets are inherently dominated by the "O" (Outside) label, as the majority of tokens in legal documents do not correspond to named entities. Consequently, a naive classifier that predicts all tokens as non-entities would achieve deceptively high accuracy while failing to identify meaningful entity information. Imbalanced learning scenarios require class-sensitive evaluation metrics beyond accuracy [11]. Therefore, a more reliable assessment of model effectiveness is obtained through precision, recall, and F1-score. Precision measures the proportion of predicted entities that are correctly classified, while recall evaluates the model's ability to identify actual entity instances within the dataset. The F1-score provides a harmonic balance between precision and recall, offering a more informative metric for imbalanced classification tasks. In this study, the balanced F1-score of 0.92 indicates that the model maintains strong detection capability across entity categories despite class imbalance. The disparity between overall accuracy and entity-specific metrics underscores the importance of class-sensitive evaluation strategies in legal NLP. Visualization of tag distributions typically reveals a dominant peak corresponding to the "O" class, confirming the skewed data distribution. Consequently, improvements in precision and recall for minority entity classes are more indicative of practical utility than marginal gains in overall accuracy. Emphasizing entity-level metrics ensures that the system effectively captures legally significant information rather than merely optimizing for majority-class prediction.This analytical perspective reinforces the necessity of robust evaluation frameworks when assessing NER systems in domain-specific and imbalanced environments such as legal text analysis.

Conclusion

This study demonstrates the effectiveness of a fine-tuned Legal-BERT-based framework for automated Named Entity Recognition in legal documents. By leveraging domain-adapted transformer representations and integrating a semantic similarity filtering mechanism, the proposed system accurately extracts structured entities—including organizations, persons, geopolitical locations, and temporal expressions—from complex legal texts. The framework addresses critical challenges inherent in legal NLP, such as domain-specific terminology,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

syntactic complexity, and class imbalance, thereby enhancing entity boundary precision and contextual consistency. Experimental evaluation confirms the robustness of the approach, achieving a precision, recall, and F1-score of 0.92, along with an overall accuracy of 0.99. While accuracy is influenced by the predominance of non-entity tokens, the balanced F1-score reflects reliable entity-level performance across categories. The deployment of the model through a Streamlit-based web application further demonstrates its practical applicability. The system enables legal professionals to upload documents, visualize extracted entities through color-coded highlighting, and access structured tabular and statistical summaries, thereby streamlining document analysis and significantly reducing manual review effort. Overall, the proposed framework establishes a scalable, robust, and application-oriented solution for automated legal entity recognition. By bridging advanced transformer-based modeling with an accessible deployment interface, this work contributes a practical tool for legal informatics and lays the groundwork for future advancements in intelligent legal text analytics.

References

[1]. A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[2]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[3]. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," in Findings of EMNLP, 2020, pp. 2898–2904.
[4]. E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in Proc. CoNLL, 2003, pp. 142–147.
[5]. E. Quevedo et al., "Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study," IEEE Access, vol. 12, pp. 145286–145317, 2024.
[6]. I. Chalkidis and I. Androutsopoulos, "A Deep Learning Approach to Contract Element Extraction," in Proc. ICAIL, 2017.
[7]. M. Peters et al., "Deep contextualized word representations," in Proc. NAACL-HLT, 2018, pp. 2227–2237.
[8]. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.
[9]. M. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv:2004.05150, 2020.
[10]. M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," in Proc. NeurIPS, 2020.
[11]. H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[12]. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. EMNLP: System Demonstrations, 2020.
[13]. D. Hendrycks et al., "Using Pre-Training Can Improve Model Robustness and Uncertainty," in Proc. ICML, 2019.
[14]. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2018.
[15]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[16]. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in Proc. NAACL-HLT, 2016, pp. 260–270.
[17]. X. Ma and E. Hovy, "End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in Proc. ACL, 2016, pp. 1064–1074.
[18]. A. Radford et al., "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.
[19]. A. Conneau et al., "Unsupervised Cross-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1183-1194
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0167

lingual Representation Learning at Scale," in Proc. ACL, 2020.

[20]. B. Settles, "Active Learning Literature Survey," University of Wisconsin-Madison, 2009.