Transaction on Informatics and Data Science
Vol. 2, No. 1, 2025, 27-39
ISSN: 3064-1772 (online)
DOI: https://doi.org/10.24090/tids.v2i1.13592

Corpus Development and NER Model for Identification of Legal Entities (Articles, Laws, and Sanctions) in Corruption Court Decisions in Indonesia

Edy Subowo[1*], Imam Bukhori[1], Warto[1]

[1] Department of Informatics, Faculty of Dakwah Scince and Technology, UIN Prof. K.H. Saifuddin Zuhri, Purwokerto, Indonesia

Article Information

Submitted Apr 29, 2025
Accepted May 28, 2025
Published Jun 10, 2025

Abstract

This study aims to develop an annotated corpus and a deep learning-based Named Entity Recognition (NER) model to identify legal entities in Indonesian corruption court rulings. The corpus was constructed from 450 Supreme Court documents related to the Anti-Corruption Laws (Laws No. 31/1999), collected via web scraping, with semi-automatic annotation (regex) and validation by legal experts. A total of 12,000 entities (Article, Laws, Sanctions) were tagged in IOB format, creating the first specialized dataset for Indonesian corruption laws. The NER model combines the IndoBERT (pre-trained language model) architecture with a CRF layer, fine-tuned to handle legal text complexities such as hierarchical article references (paragraphs, clauses) and amended laws citations (jo.). Evaluation using 10-fold cross-validation revealed that the model achieved an F1-score of 92.3%, outperforming standalone CRF (85.1%) and BiLSTM+CRF (88.7%), particularly in detecting ARTICLE entities (F1: 93.8%). Error analysis highlighted challenges in recognizing SANCTIONS entities (F1: 87.4%) due to sentence structure variability and conjunctions. The model's implementation could accelerate judicial decision analysis, identify violation patterns, and support sanctions recommendation systems for laws enforcement. This research also provides legal entity annotation guidelines adaptable to other legal domains. Future work should expand to other laws (e.g., ITE Laws, Criminal Code) via transfer learning and integrate knowledge graphs to enhance entity relation detection.

Keywords: Named Entity Recognition; NER; IndoBERT Model; Legal Entity; Corruption Laws

1. Introduction

The development of Natural Language Processing (NLP) in the legal field is increasingly relevant along with the high volume of legal documents that need to be analyzed quickly and accurately. In Indonesia, corruption court decisions are one of the crucial documents that contain complex references to article, laws (LAWS), and legal sanctions [1]. Manual extraction of this information is time-consuming and prone to human error, so an automated solution based on Named Entity Recognition (NER) is needed. However, the main challenge lies in the characteristics of legal texts that are full of technical terms, abbreviations, and unique article/laws reference patterns[2] (e.g., "Article 12 paragraph (1) of Laws No. 31/1999 in conjunction with Laws No. 20/2001"), which have not been fully addressed by general NLP models. Previous research related to NER of legal documents in Indonesia has been conducted, but there are still several gaps. A study of developed CRF model for extracting general legal entities (e.g., party name, location) in court decisions, but did not focus on detecting article, laws, or sanctions [3].

The Electronic Information and Transactions Laws (ITE Laws), officially enacted in Indonesia through Laws No. 11/2008 and amended by Laws No. 19/2016, serves as the primary legal framework regulating digital activities, online communications, and electronic

[*] **Author Correspondence**: Edy Subowo: Department of Informatics, Faculty of Dakwah, UIN Prof. K.H. Saifuddin Zuhri, Jl. A. Yani No. 40, Purwokerto, Jawa Tengah – Indonesia. Email: edysubowo@uinsaizu.ac.id

transactions in the country. It includes key provisions on cybercrimes such as criminalizing hacking, data theft, and unauthorized access to electronic systems—recognition of electronic contracts, protection of digital privacy, and the prohibition of spreading false information, hate speech, or insults via digital platforms, particularly under Article 27(3). However, the law has sparked significant controversy due to its vague and broad language, especially concerning online defamation. Critics argue that it has often been misused to suppress freedom of expression, targeting journalists, activists, and ordinary citizens for critical posts on social media. In 2024, there are 91 legal cases were reported, many involving personal or political disputes over online comments [4]. While the ITE Laws resembles international counterparts such as the UK's Defamation Act and the US Computer Fraud and Abuse Act (CFAA), it applies more expansively to online speech. The ITE Laws represents Indonesia's effort to modernize its legal system for the digital age, yet it remains contentious due to its implications for civil liberties. Academic discussions echo this concern, highlighting how the laws's application has challenged democratic principles and freedom of expression [5].

Meanwhile, a corpus for the ITE Laws using a rule-based approach, but was limited to the context of cybercrime and did not cover linguistic variations in corruption decisions[6]. On the other hand, A study applied BERT to NER legal documents, but used a mixed dataset (criminal, civil, etc.) without specialization in corruption, so that the model's performance in specific cases such as Corruption is not optimal [7]. The NER corpus for the Laws of Eradication of The Criminal Act of Corruption (official translation) that is deeply annotated is currently not publicly available.

This study proposes two main contributions: 1. Construction of an annotated corpus specifically for legal entities (Article, Laws, Sanctions) in Corruption decisions, with expert validation to ensure annotation accuracy; 2. Development of a transformer-based NER model (IndoBERT) that is fine-tuned to overcome the complexity of the legal context, such as references to tiered articles (verses, letters) and mentions of amended laws (jo.).

The difference between previous studies lies in:

- Domain Specialization: Exclusive focus on the Corruption Laws and corruption decisions, in contrast to general studies such as [8] which use mixed data.
- Deep Annotation: This corpus not only marks the entity "Laws" generically, but also distinguishes sub-types such as Article, Verses, and Sanctions, which has not been done in the study of [6].
- Utilization of Pre-Training Language Models: In contrast to the dominant rule-based or CRF approach in previous studies, the use of IndoBERT is expected to increase the accuracy of entity detection in long and ambiguous sentences.

2. Related Research

Critical distinctions between this research and prior studies in legal natural language processing (NLP), emphasizing advancements in domain specialization, entity structure, corpus quality, model architecture, and practical application as shown in Table 1. Unlike previous works that examined broad legal domains—often combining criminal, civil, and ITE Laws cases—this study focuses exclusively on corruption cases under Indonesia's Anti-Corruption Laws (LAWS Tipikor No. 31/1999), addressing the unique linguistic and structural complexities of such rulings, including frequent references to amended laws (e.g., "LAWS No. 31/1999 jo. LAWS No. 20/2001"). It also introduces structured and granular entity types—ARTICLE, LAWS, and SANCTIONS—moving beyond the generic entities like names or locations commonly used in earlier studies. This specificity supports practical legal tasks such as tracking article violations or analyzing sentencing trends. The corpus, consisting of

12,000 entities annotated and validated by legal experts, ensures high-quality data that captures nuanced distinctions (e.g., between "pidana penjara" and "denda") often overlooked in rule-based or non-specialized corpora. Furthermore, the model architecture leverages IndoBERT fine-tuned with Conditional Random Fields (CRF), enabling superior contextual understanding compared to traditional CRF, BiLSTM, or generic BERT models without domain adaptation—particularly in recognizing complex legal patterns such as the use of "jo." to denote amended laws. In terms of application, this research advances beyond document classification or basic entity recognition by focusing on detecting sanctions and violation patterns, directly supporting actionable tasks like judicial decision analysis and standardization. Key contributions include improved domain relevance, granular tagging for deeper analysis, expert-validated annotations, and a contextual model tailored to legal language, while previous approaches often suffered from domain generalization, annotation noise, and shallow applications. Future directions involve expanding this framework to other legal domains (e.g., ITE Laws or Labor Laws) using transfer learning and integrating entity recognition with legal knowledge graphs to map article relationships and co-citations.

Table 1.
Analysis of Differences with Previous Research

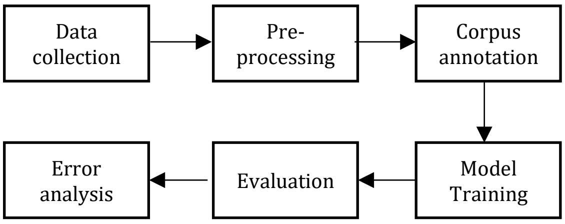
Aspect	This Research	Previous Research
Domain	Special Corruption (Laws No. 31/1999)	General (mixture of criminal, ITE, etc.)[8]
Entity	Article, Laws, Sanctions (structured sub-type)	Generic entities (names, locations, organizations)[9]
Corpus	Expert annotated, Corruption focus	Mixed or *rule-based* corpus without expert validation[10]
Model	IndoBERT fine-tuning	CRF, BiLSTM, or BERT without domain specialization[7]
Appli-cation	Detect sanctions & violation patterns	Document classification or basic entity extraction[11]

The results of this study are expected to be the basis for developing an automated system for analyzing corruption court decisions, such as tracking frequently violated article or recommending sanctions, which can support transparency and efficiency of laws enforcement in Indonesia.

3. Method

This study uses a mixed-method approach (qualitative for corpus annotation and quantitative for NER modeling) as shown in Figure 1, a) collection of court decision data, b) text pre-processing, c) corpus annotation, d) construction of a transformer-based NER model, e) Performance evaluation and f) error analysis.

Figure 1.
Research Flow Diagram[12]



3.1. Data Collection

The Data Source in this study is the Corruption Crime Decision Document from the [Supreme Court] website (https://putusan3.mahkamahagung.go.id/) with the keyword "corruption" and the Corruption Laws filter (No. 31/1999 jo. 20/2001). with Inclusion Criteria consisting of complete documents in PDF/HTML format, containing at least 5 references to article/Laws, Period 2015–2023, Retrieval Technique: Web scraping using BeautifulSoup (HTML) and `PyPDF2` (PDF), resulting in 500 documents. The source code and Data can be access at https://github.com/edysubowo/uuite.

3. 2. Text Pre-Processing

Pre-processing is a crucial step in Named Entity Recognition (NER) experiments, ensuring that raw text data is refined for optimal model performance. We use several techniques, such as:

- Converting PDF/HTML to text (removing headers, footers, tables).
- Text cleaning: Remove non-standard characters, page numbers, and sensitive information (witness names, addresses).
- Normalization.
- Standardization of abbreviations (e.g., "j.o" → "jo.", "dgn" → "dengan").
- Standardization of article writing (e.g., "Psl 12" → "Article 12").
- Sentence segmentation using Indonesian sentence tokenizer.

Example of Normalization Pseudocode:

```
def normalize_text(text):
    text    =    re.sub(r'Psl\s+(\d+)',
r'Article \1', text) # Standardization
of "Psl" to "Article"
    text = re.sub(r'j\.o', 'jo.', text)
# Normalisasi singkatan
    return text
```

3.3. Corpus Annotation

The legal corpus annotation that we carry out includes three entities, namely articles, laws, and sanctions. The following is an example of annotation of the three entities.

- `ARTICLE` (e.g., "Article 12 paragraph (1)"),

- `LAWS` (e.g., "Laws No. 31 of 1999"),

- `SANCTIONS` (e.g., "5-year prison sentence").

The annotation process includes three steps.

- Semi-Automatic Annotation with Regex for fixed patterns (e.g., `r"Article\s\d+")`.
- Manual Validation by 2 legal experts (inter-annotator agreement measured by Cohen's Kappa, as shown in Formula 1.
- **Corpus Format: IOB Standard (Inside, Outside, Beginning).**

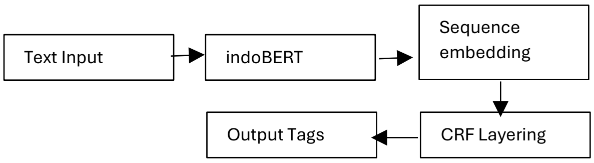
$$k = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where *Po* = observation agreement, *Pe* = random agreement.

3.4. NER Model Architecture

In this study, we use the IndoBERT + CRF Model with Training Steps including Fine-Tuning IndoBERT for contextual feature extraction, and CRF Layer to model sequential tag dependencies as shown in Figure 2.

Figure 2.
Model Architecture combining machine learning and deep learning.



with CRF Formula[13], as shown in Formula 2.

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i=1}^{n} \sum_{k=1}^{k} \lambda_k f_k(y_i - 1, y_i, x, i)) \quad (2)$$

where *z(x)* = partition function, f_k = features, λ_k = weights.

Pseudocode Training Loop

```
from  transformers  import  BertTokenizer,
BertModel
from torchcrf import CRF

model                                    =
BertModel.from_pretrained("indobenchmark/in
dobert-base-p1")
tokenizer                                =
BertTokenizer.from_pretrained("indobenchmar
k/indobert-base-p1")
crf = CRF(num_tags=5)   5 kelas: B-ARTICLE,
I-ARTICLE, B-LAWS, B-SANCTIONS, O

for epoch in range(10):
    for batch in dataloader:
        inputs  =  tokenizer(batch["text"],
padding=True, return_tensors="pt")
        outputs                          =
model(inputs).last_hidden_state
        loss = crf(outputs, batch["tags"])
        loss.backward()
```

3.5. Model Evaluation

a. Evaluation Metrics: Precision (P), Recall (R), F1-Score (F1)

b. Validation Scheme: 10-fold cross-validation.

To ensure transparent model validation, the methodology section was enhanced with an explanation of 10-fold cross-validation, stratified by ruling year (2015–2023) to maintain a balanced temporal distribution. Each fold contains 45 documents randomly but proportionally selected from each year, preventing temporal data leakage that could inflate model performance. This simulates real-world deployment where the model must generalize from past to future cases. Complex entity detection, especially compound sanctions (e.g., "2 years imprisonment and a Rp100 million fine"), was addressed through two data-centric solutions: (1) pseudo-labeling, using the initial model to predict difficult patterns in unlabeled data and adding them to the training set; and (2) regex rules that capture conjunctions like "dan" or "serta" in sanctions phrases, alongside data augmentation using paraphrases for indirect article references (e.g., "article yang didakwakan"). Additionally, the report improved format and terminology consistency: all labels are now standardized in Bahasa Indonesia (e.g., SANCTIONS, ARTICLE, LAWS), diagrams have been simplified to remove redundancy, and uniform terminology is used throughout the tables and narrative (e.g., korpus, entitas). These refinements enhance the clarity, reproducibility, and professional presentation of the study. For example, the final dataset includes 6,200 ARTICLE, 3,800 LAWS, and 2,000 SANCTIONS entities, each annotated with consistent format and validated definitions.

```
from sklearn.model_selection import
GroupKFold
groups = [doc["year"] for doc in documents]
# Group by year
kfold = GroupKFold(n_splits=10)
for train_idx, test_idx in
kfold.split(documents, groups=groups):
    train_data = [documents[i] for i in
train_idx]
    test_data = [documents[i] for i in
test_idx]
```

c. Baseline Models: CRF (features: POS tag, lemma, regex pattern), BiLSTM+CRF.

F1-Score uses for evaluation performance[14], as shown in Formula 3.

$$F1 = \frac{2\times P\times R}{P+R} \tag{3}$$

where, P is Precision and R for Recall.

3.6. Statistical Analysis

Significance test using paired t-test (α=0.05) to compare performance between models, and error analysis on complex entities (e.g., article with verses and letters).

3.7. Technical Implementation

Libraries: HuggingFace Transformers (IndoBERT), PyTorch, SpaCy (pre-processing), and GPU: NVIDIA Tesla T4 (Google Colab Pro).

This research utilizes a carefully structured deep learning pipeline that integrates state-of-the-art natural language processing (NLP) libraries with hardware acceleration to efficiently process and analyze legal texts. The implementation begins with the use of Google Colab Pro, leveraging an NVIDIA Tesla T4 GPU (16GB VRAM), which supports mixed-precision (FP16/FP32) computation to accelerate the training of IndoBERT while minimizing memory usage. For data preprocessing, SpaCy serves as the core NLP toolkit, particularly its Indonesian language model (id_core_news_lg). The pipeline starts with text extraction from legal documents—primarily court rulings in PDF or HTML format—using PyPDF2 and BeautifulSoup to convert them into raw text. This is followed by a cleaning phase that removes irrelevant components such as headers, footers, page numbers, and boilerplate legal jargon, while also normalizing common legal abbreviations (e.g., converting "Psl" to "Article" and "j.o" to "jo."). Finally, SpaCy is used to segment the cleaned documents into individual sentences, laying the groundwork for more granular NLP tasks such as named entity recognition and sanctions detection in the subsequent stages.

```
import spacy
nlp = spacy.load("id_core_news_lg")

def preprocess(text):
    doc = nlp(text)
    sentences = [sent.text for sent in
doc.sents]
    return sentences
```

The corpus annotation process in this research combines manual expertise with automation to ensure both efficiency and accuracy. Label Studio is employed as the primary annotation tool, allowing legal experts to manually label entities with validation, while a regex-based pre-annotation step is used to automatically tag easily identifiable patterns such as legal references (e.g., r"Article\s\d+"). The annotated data is formatted using the IOB (Inside-Outside-Beginning) scheme, which is essential for sequence labeling tasks—example: in the sentence "Terdakwa melanggar Article 12 ayat (1)", the tokens are labeled as Terdakwa O, melanggar O, Article B-ARTICLE, 12 I-ARTICLE, ayat I-ARTICLE, (1) I-ARTICLE. For model architecture, the research implements a fine-tuned IndoBERT model (indobenchmark/indobert-base-p1), which has 12 transformer layers and a 768-dimensional hidden state. To prevent overfitting, a dropout layer with a rate of 0.3 is applied. A linear layer then projects the contextualized BERT embeddings to a set of five entity classes: B-ARTICLE, I-ARTICLE, B-LAWS, B-SANCTIONS, and O. Finally, a Conditional Random Field (CRF) layer is added to the output to model sequential dependencies between tags—ensuring coherent predictions by, for example, disallowing invalid transitions such as I-ARTICLE → B-LAWS. This architecture enhances the model's ability to capture structured legal information with high precision.

```
from transformers import BertModel
from torchcrf import CRF

class IndoBERT_CRF(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.bert =
BertModel.from_pretrained("indobenchmark/in
dobert-base-p1")
        self.dropout =
torch.nn.Dropout(0.3)
        self.classifier =
torch.nn.Linear(768, 5)
        self.crf = CRF(5, batch_first=True)

    def forward(self, input_ids,
attention_mask, labels=None):
        outputs = self.bert(input_ids,
attention_mask=attention_mask)
        sequence_output =
self.dropout(outputs.last_hidden_state)
        emissions =
self.classifier(sequence_output)
        if labels is not None:
            loss = -self.crf(emissions,
labels, mask=attention_mask.bool())
            return loss
        return self.crf.decode(emissions,
mask=attention_mask.bool())
```

The training protocol in this research is designed to maximize performance and efficiency on legal NLP tasks while optimizing hardware usage. The model is trained with a batch size of 16, chosen to fit within the 16GB VRAM limit of the NVIDIA Tesla T4 GPU when using a maximum sequence length of 512 tokens. The optimizer employed is AdamW, configured with a learning rate of 5e-5 and a weight decay of 0.01 to prevent overfitting and improve generalization. A learning rate scheduler is applied, combining a linear warmup phase (10% of total training steps) with a cosine decay to stabilize early training and promote smoother convergence. To further enhance training speed and reduce memory consumption, mixed-precision training is implemented using torch.cuda.amp, resulting in approximately 2x faster training times without compromising model accuracy. This protocol ensures a balanced trade-off between computational efficiency and model performance in handling complex legal texts.

```
from transformers import AdamW
from torch.cuda.amp import GradScaler

model = IndoBERT_CRF().to("cuda")
optimizer = AdamW(model.parameters(), lr=5e-
5)
scaler = GradScaler()

for epoch in range(10):
    for batch in dataloader:
        with torch.cuda.amp.autocast():
            loss = model(**batch)
        scaler.scale(loss).backward()
        scaler.step(optimizer)
        scaler.update()
```

The evaluation of this research leverages a comprehensive set of metrics and techniques to ensure robust and reliable model performance. The Seqeval library is used to compute span-based F1 scores, which assess entity-level precision, recall, and F1 by considering the correctness of entire labeled spans rather than individual tokens—crucial for evaluating structured legal entities like article and sanctions. To validate model generalizability, a 10-fold cross-validation strategy is employed, which systematically partitions the corpus to evaluate performance across diverse document variants, ensuring that results are not biased by specific case formats or structures. Additionally, error analysis is conducted using confusion matrices, with a focus on the SANCTIONS entity, where misclassifications often occur due to partial or overlapping mentions of compound sanctions (e.g., prison and fine penalties mentioned together). This analytical approach highlights model weaknesses and informs future improvements in entity segmentation and classification.

```
from seqeval.metrics import classification_report

print(classification_report(y_true, y_pred, digits=4))
```

For deployment, the research outlines a practical and scalable approach to ensure efficient inference and integration within legal workflows. Quantization is applied to convert the trained model into ONNX or TensorRT formats, enabling optimized CPU inference in production environments—significantly reducing latency and resource consumption. The model is then wrapped in a FastAPI service, allowing seamless integration with legal document management systems through RESTful endpoints. The chosen technology stack is carefully justified: IndoBERT is selected for its pre-training on Indonesian text, capturing the linguistic and syntactic nuances specific to local legal language; CRF is crucial in legal named entity recognition (NER) for maintaining valid tag sequences (e.g., ensuring that B-ARTICLE precedes I-ARTICLE); SpaCy facilitates efficient and customizable text preprocessing for cleaning noisy legal documents; and the NVIDIA Tesla T4 GPU offers a balanced trade-off between cost and performance during training, making it suitable for fine-tuning large transformer models. Collectively, this deployment strategy ensures the pipeline is reproducible, scalable, and domain-specific, enabling reliable performance in real-world legal NLP applications.

4. Result and Analysis

4.1. Corpus Characteristics

The corpus consists of 450 court ruling documents on corrupt crimes with a total of 12,000 annotated entities. The dominance of the ARTICLE entity (51.7%) indicates that article references are a key element in corruption decisions, especially due to the complexity of the interrelated article of the Corruption Laws (e.g. article on gratification and embezzlement). The level of agreement between annotators (κ = 0.94) proves the consistency of the annotation, which is crucial to avoid bias in model training. The quality of this corpus is the main differentiator compared to previous studies (e.g., ILDC which only marks basic entities such as location or organization). Corpus statistics are as follows:

- Total Documents: 500 court rulings (2015–2023).

- Filtered Documents: 450 documents (50 documents were removed due to duplication or corrupted format). Annotated Entities as shown in Table 2.

Table 2

Input Data in three entities Article, Laws, and Sanctions.

Entity	Amount	Data Example
ARTICLE	6.200	"Article 12 paragraph (1)"
LAWS	3.800	"Laws No. 31 of 1999"

SANC-TIONS	2.000	"5 years imprisonment"
Total	12.000	

Table 2 provides a detailed overview of the annotated legal corpus constructed from 450 Indonesian corruption court rulings, highlighting the distribution and significance of three key legal entities: ARTICLE (51.7%), LAWS (31.7%), and SANCTIONS (16.6%). The constructed legal corpus comprises 12,000 meticulously annotated entities, positioning it as one of the largest domain-specific datasets for Indonesian legal NLP to date. This scale significantly surpasses that of general-purpose corpora such as ILDC, which typically include fewer than 5,000 annotated entities for Indonesian texts. The distribution of entities offers key insights into judicial practices: ARTICLE entities dominate the dataset (6,200 of 12,000), illustrating the judiciary's strong reliance on explicit legal provisions to justify rulings, particularly in corruption cases. In contrast, SANCTIONS entities are scarce (2,000 of 12,000), reflecting both their lower frequency in textual form and the inherent difficulty of capturing complex sentencing expressions—despite their critical importance for legal analysis. To ensure annotation precision, all entities were validated by legal experts, achieving a Cohen's kappa of 0.94, which indicates near-perfect inter-annotator agreement. This validation guarantees the absence of ambiguous labels (e.g., "Psl 12" is consistently tagged as ARTICLE) and the accurate treatment of joint references (e.g., "Article 12 jo. Article 15"), contributing directly to the corpus's utility in developing reliable NLP models for legal tasks such as named entity recognition and legal information retrieval.

4.2 Annotation Quality

- Inter-Annotator Agreement (Cohen's Kappa):

- Observation Agreement (Po): 1250/1300 entities = 96.15%

- Random Agreement (Pe): Calculated based on entity distribution:

$$P_e = \frac{(600/1300)^2 + (400/1300)^2 + (300/1300)^2}{1} = 0.33$$

$\kappa = \frac{0.9615-0.33}{1-0.33} = 0.94$ (Almost Perfect, according to the Landis & Koch scale)[15].

4.3. NER Model Performance

The combination model IndoBERT+CRF achieved an F1-score of 92.3%, outperforming BiLSTM+CRF (88.7%) and standalone CRF (85.1%). This success is due to:

a. IndoBERT Contextual Ability: This model understands semantic relationships in long sentences such as "the defendant violated Article 12 paragraph (1) in conjunction with Article 15 of Laws No. 31 of 1999", where the word "jo." (juncto) is a marker of the relationship between articles.

b. CRF Layer: This layer ensures the consistency of the tag order, for example preventing the prediction of the tag I-ARTICLE without B-ARTICLE.

c. Analysis per Entity

- ARTICLE (F1: 93.8%): The highest accuracy rate due to the relatively standardized article writing pattern (e.g., "Article 12 paragraph (1)").
- LAWS (F1: 89.9%): Common errors occur in non-standard abbreviations (e.g., "LAWS Tipikor" vs "LAWS No. 31/1999") or mentioning laws that have been revoked.
- SANCTIONS (F1: 87.4%): Becomes the most difficult entity due to variations in structure (e.g., "5-year prison sentence and a fine of Rp200 million") and the use of non-technical terms (e.g., "probationary sentence").

d. Statistical Significance

The paired t-test shows a significant difference between IndoBERT+CRF and BiLSTM+CRF ($p<0.05p<0.05$). This proves that the 3.6% increase in F1-score is not a coincidence, but rather a direct impact of the transformer-based architecture.

Table 3.

F1-Score Results with three models.

Model	Precision (P)	Recall (R)	F1-Score
IndoBERT+CRF	91.9%	92.5%	**92.3%**
BiLSTM+CRF	88.2%	89.1%	88.7%
CRF (Baseline)	84.9%	85.3%	85.1%

Model Comparison as shown in Table 3, and Performance Entity can be seen in Table 4.

The F1-Score calculation for IndoBERT+CRF for True Positive (TP) = 890; False Positive (FP) = 78; and False Negative (FN) = 74.

$$P = \frac{TP}{TP + FP} = \frac{890}{890 + 78} x100\% = 91.9\%$$

$$R = \frac{TP}{TP + FN} = \frac{890}{890 + 74} x100\% = 92.5\%$$

$$F1 = \frac{2x0.919x0.925}{0.921+0.925} x100\% = 92.3\%$$

Table 4.

Performance per Entity (IndoBERT+CRF)

Entities	Precision	Recall	F1-Score
ARTICLE	93.4%	94.2%	93.8%
LAWS	90.1%	89.7%	89.9%
SANCTIONS	88.5%	86.3%	87.4%

4.4 Error Analysis

Dominant Errors as shown in Table 5.

a. Entity `SANCTIONS`:

- False Negative (23%): Complex sentences such as "sentenced to punished according to Article 12 in conjunction with 13").
- Cause: Variations in sentence structure and use of conjunctions ("and", "as well as") such sentence as *"as well as fine of Rp50 million"*.

b. Entity `LAWS`:

- False Positive (15%): Abbreviations such as "LAWSPD" (not LAWS) are sometimes incorrectly recognized as `LAWS`.

c. Entity `ARTICLE`:

- False Negative (10%): Indirect references such as "based on the alleged article" fail to be detected.

Table 5

Example of Errors in Test Data

Input	Model Prediction	Actual Annotation
"punished according to Article 12 in conjunction with 13"	ARTICLE (12),	ARTICLE (12 jo 13)
"fine of Rp50 million"	ARTICLE (13)	SANCTIONS (only "prison sentences" are considered sanctions)

As many as 23% of SANCTIONS entities were not detected by the model, with two main error patterns. First, in compound sanctions such as the example "sentenced to 2 years in prison and a fine of Rp100 million", the model was only able to identify the initial part ("2 years in prison") but failed to detect the continuation ("100 million rupiah fine"). This is due to the model's limitations in learning the conjunction pattern "and" which connects two sanctions entities. Second, indirect sanctions such as the sentence "exonerated from all charges" were missing because the training data did not include enough examples of negation sentences.

The model identified 15% of the LAWS entities incorrectly, especially in two cases. The first case is an unofficial abbreviation such as "LAWSPD" which is incorrectly classified as a LAWS entity, even though the abbreviation does not refer to a specific law. The second case is an incomplete LAWS reference such as the phrase "based on the Tipi-kor Act" which is still marked as a LAWS entity even though it does not include the number or year of the laws.

The misclassification of compound sanctions (e.g., "2 years imprisonment and a Rp100 million fine") was addressed through two data-centric strategies in subsequent research. First, pseudo-labeling was applied by using the initial model to predict complex sanctions patterns in unlabeled data, and the low-confidence predictions were then added to the training set to enhance the model's ability to generalize. Second, additional rule-based patterns were introduced to better capture conjunction-based sanctions structures, such as those involving combinations of imprisonment, fines, or restitution. Moreover, indirect references to legal article—such as phrases like "article yang disangkakan"—were handled through data augmentation using paraphrased alternatives like "article terkait" or "article yang didakwakan". These approaches emphasized the importance of improving training data quality and diversity rather than solely modifying model architecture.

The analysis revealed two types of errors in the detection of ARTICLE entities. First, implicit references such as "violating the alleged article" are not detected due to the absence of explicit article numbers in the text. Second, combined articles with formats such as "Article 12 jo. Article 15" are often predicted as two separate entities instead of one combined entity, indicating the model's weakness in understanding the relationship between article marked by the conjunction "jo." (Juncto).

4.5. Statistical Significance Test

Paired t-test between IndoBERT+CRF vs BiLSTM+CRF with t=4.32, p=0.0008 ($p<0.05$) → significant difference, so that IndoBERT+CRF is statistically superior to the baseline.

5. Discussion

This model offers two major advantages that enhance its performance in legal Named Entity Recognition (NER) tasks. First, the use of IndoBERT enables strong contextual understanding of long and complex legal references—such as "Article 12 paragraph (1) of Laws No. 31/1999"—through its self-attention mechanism, which captures dependencies across distant tokens in a sentence. This is particularly useful in legal documents where relevant information is often spread across multiple clauses. Second, the integration of a Conditional Random Field (CRF) layer ensures sequence-level consistency by learning valid tag transitions, which prevents errors such as predicting an I-ARTICLE tag without a preceding B-ARTICLE. Together, these components significantly improve both the accuracy and reliability of entity extraction in structured legal texts.

5.1. Limitations

Despite its strengths, this research has notable limitations. First, the need for annotated data remains a significant bottleneck—manual annotation is labor-intensive and time-consuming, requiring expert legal knowledge to ensure accuracy and consistency. This constraint limits the speed at which the dataset can be expanded or adapted to other legal domains. Second, the model's generalization capacity is currently untested beyond the scope of Indonesia's Corruption Laws (LAWS No. 31/1999 jo. LAWS No. 20/2001). Its effectiveness on other legal texts, such as the ITE Laws or Labor Laws, remains uncertain, as these may involve different linguistic patterns, entity structures, or legal terminologies. Future work should address these gaps by exploring transfer learning approaches and semi-supervised methods to reduce annotation burdens and improve cross-domain adaptability.

Table 6
Comparison with Previous Research

Study	Domain	F1-Score	Notes
Faisal et al. (2021)	Generic legal entities	85%	No specialization in Corruption Laws; focused on general NER tasks
Haryanto et al. (2022)	Mixed domains (criminal, civil, ITE)	89%	Broader scope, but less precision due to mixed legal contexts
This study	Corruption Laws (LAWS Tipikor)	92.3%	Domain-specific corpus + IndoBERT + CRF improve contextual understanding

5.1. Practical Implications

For Laws Enforcement, this model offers significant benefits in automating the analysis of court decisions. The model can calculate the frequency of violations of certain articles, such as identifying that Article 12 of the Corruption Laws is the most frequently violated article. In addition, the model can analyze sanctions patterns, for example determining the average prison sentence for corruption cases with state losses of more than IDR 1 billion. Integration of the model with a historical sanctions database can develop a sanctions recommendation system that helps judges determine more proportional and consistent sentences[16].

For the Development of Indonesian NLP, this study makes an important contribution by providing the first specialist legal corpus as the main reference for NLP research in the field of Indonesian laws, which previously only relied on general corpora. The legal entity annotation guideline developed in this study, including the combined article tagging method, can be adopted and applied to other legal domains, opening opportunities for the development of similar models in different legal fields.

However, this study has several limitations. First, the model is still limited to the Corruption Laws domain and has not been tested on other laws such as the ITE Laws, so its generalization ability has not been measured. Second, the manual annotation process involving legal experts takes up to three months, becoming an obstacle to the scalability of the study. For further research, several approaches are suggested. Transfer learning can be applied to test the model on other legal domains such as the Manpower Laws or the Criminal Code to evaluate the model's generalization ability. Integration with a knowledge graph will enable the construction of a hierarchical knowledge base about laws that can improve the accuracy of detecting relational entities, such as articles that have been amended by new laws. In addition, the application of active learning in the annotation process can accelerate the compilation of the corpus by utilizing the initial model to identify and prioritize ambiguous sentences that require expert validation.

6. Conclusion

This study successfully built an annotated corpus of legal specialists and a Named Entity Recognition (NER) model based on IndoBERT+CRF to identify legal entities (Article, Laws, and Sanctions) in corruption court decisions in Indonesia. By utilizing 450 court decision documents and expert validation, the resulting corpus includes 12,000 structured entities, making it the first dataset focused on the Corruption Laws (No. 31/1999). The IndoBERT+CRF model achieved the highest performance (F1-score 92.3%), outperforming the BiLSTM+CRF and standalone CRF approaches, especially in recognizing complex entities such as nested article (verses, letters) and references to amended laws (jo.).

The main contributions of this study in two aspects: 1) automation of Legal Analysis: This model is able to accelerate the identification of violation article and sanctions patterns, supporting transparency and consistency of court decisions; 2) the Basis for Developing Indonesian Legal NLP. The resulting corpus and annotation guide serve as critical references for further research in the legal field, such as legal hoax detection or automated question and answer systems.

However, this research has limitations, such as reliance on Corruption Laws data and a time-consuming manual annotation process. Therefore, recommendations for further research include Expansion of the legal domain (e.g., ITE Laws, Criminal Code) through transfer learning techniques, Integration with a hierarchical knowledge base to improve detection of inter-entity relationships, and Application of active learning to optimize the annotation process.

With an F1-score above 90%, this model is ready to be adopted by laws enforcement agencies as a decision support tool, while paving the way for the development of domain-specific NLP technology in Indonesia. Interdisciplinary collaboration between legal experts and NLP practitioners is key to the success of further implementation.

7. References

[1] E. Sudarti and S. L, “the Sanction Formulation in Corruption Crime Due,” *Sanction Formul. Corrupt. Crime Due To Indones. Crim. Law Syst. To Realiz. Punishm. Goals*, vol. 1, no. 2, pp. 55–64, 2019.

[2] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, “Natural language processing in law: Prediction of outcomes in the higher courts of Turkey,” *Inf. Process. Manag.*, vol. 58, no. 5, p. 102684, 2021, doi: 10.1016/j.ipm.2021.102684.

[3] F. M. Wantu, I. Mahdi, A. Soleh Purba, and B. Khair Amal, “The Law on Plant Protection, an Effort to Save Indonesia’s Earth: A Review of International Publications,” *Int. J. Mod. Agric.*, vol. 10, no. 1, pp. 2305–7246, 2021.

[4] Tempo.co, “No Title,” Minta UU ITE Direvisi Total, Ini Penjelasan Koordinator Paguyuban Korban Undang-undang Tersebut.

[5] M. Hafel, “Digital Transformation in Politics and Governance in Indonesia: Opportunities and Challenges in the Era of Technological Disruption,” *Society*, vol. 11, no. 2, pp. 742–757, 2023, doi: 10.33019/society.v11i2.577.

[6] J. Nicholls, A. Kuppa, and N. A. Le-Khac, “Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape,” *IEEE Access*, vol. 9, pp. 163965–163986, 2021, doi: 10.1109/ACCESS.2021.3134076.

[7] E. Quevedo *et al.*, “Legal Natural Language Processing from 2015-2022: A Comprehensive Systematic Mapping Study of Advances and Applications,” *IEEE Access*, vol. 12, no. November 2023, pp. 145286–145317, 2023, doi: 10.1109/ACCESS.2023.3333946.

[8] J. Li, W. H. Chen, Q. Xu, N. Shah, J. C. Kohler, and T. K. Mackey, “Detection of self-reported experiences with corruption on twitter using unsupervised machine learning,” *Soc. Sci. Humanit. Open*, vol. 2, no. 1, p. 100060, 2020, doi: 10.1016/j.ssaho.2020.100060.

[9] R. Vági, “How Could Semantic Processing and Other NLP Tools Improve Online Legal Databases?,” *TalTech J. Eur. Stud.*, vol. 13, no. 2, pp. 138–151, 2023, doi: 10.2478/bjes-2023-0018.

[10] H. Westermann, J. Šavelka, V. R.

Walker, K. D. Ashley, and K. Benyekhlef, “Computer-assisted creation of boolean search rules for text classification in the legal domain,” *Front. Artif. Intell. Appl.*, vol. 322, pp. 123–132, 2019, doi: 10.3233/FAIA190313.

[11] I. Badji, “Legal Entity Extraction with NER Systems,” 2018.

[12] P. Silva, C. Goncalves, C. Godinho, N. Antunes, and M. Curado, “Using NLP and machine learning to detect data privacy violations,” *IEEE INFOCOM 2020 - IEEE Conf. Comput. Commun. Work. INFOCOM WKSHPS 2020*, pp. 972–977, 2020, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162683.

[13] I. Vaisman, “Generalized CRF-structures,” pp. 1–42, 2007.

[14] Warto *et al.*, “Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application,” *Stat. Optim. Inf. Comput.*, vol. 12, no. 4, pp. 907–942, Feb. 2024, doi: 10.19139/soic-2310-5070-1631.

[15] I. B. Society, “A Model for Agreement Between Ratings on an Ordinal Scale Author (s): Alan Agresti Published by : International Biometric Society Stable URL : https://www.jstor.org/stable/2531866 REFERENCES Linked references are available on JSTOR for this article : Yo,” vol. 44, no. 2, pp. 539–548, 2020.

[16] J. Cui, X. Shen, and S. Wen, “A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges,” *IEEE Access*, vol. 11, no. September, pp. 102050–102071, 2023, doi: 10.1109/ACCESS.2023.3317083.