IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 5, May, 2025, pp. 1627-1633
Journal homepage: http://iieta.org/journals/mmep

Learning-Based Information Extraction to Obtain Prominent Named Entities in Indonesian Court Decision Documents

Firdaus Solihin[1,2*], Indra Budi[1], Eka M.S. Rochman[2], Fifin A. Mufarroha[2], Ahmad A. Ramdlany[3], Deshinta A. Dewi[4]

[1] Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia
[2] Faculty of Engineering, University of Trunojoyo Madura, Bangkalan 69162, Indonesia
[3] Faculty of Law, University of Trunojoyo Madura, Bangkalan 69162, Indonesia
[4] Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia

Corresponding Author Email: fsolihin@trunojoyo.ac.id

https://doi.org/10.18280/mmep.120517

Received: 28 December 2024
Revised: 18 April 2025
Accepted: 24 April 2025
Available online: 31 May 2025

Keywords:
information extraction, legal named entity recognition, legal NER, BERT, court decision

ABSTRACT

The increasing number of cases processed in Indonesian courts has led to a rapid growth in court decision documents, which contain crucial legal information. However, due to their unstructured textual nature, diverse classifications, linguistic variations, and inconsistent document structures, extracting meaningful information from these documents remains a significant challenge. This study presents a comparative analysis of machine learning approaches for information extraction (IE) from Indonesian court decisions in criminal tribunal, employing Conditional Random Fields (CRF), Support Vector Machines (SVM), Bidirectional Long Short-Term Memory (Bi-LSTM), and Bidirectional Encoder Representations from Transformers (BERT). Experimental results demonstrate that CRF outperforms SVM in terms of F1-score (0.65 vs. 0.19), indicating its relative robustness for structured prediction tasks. Meanwhile, Bi-LSTM achieves an accuracy of 0.37, reflecting limitations in handling the linguistic complexity of legal texts. Notably, BERT significantly surpasses all other methods, achieving an outstanding accuracy of 0.96. The superior performance of BERT is attributed to its deep contextualized representation and ability to leverage pre-trained knowledge, making it highly effective for handling domain-specific variability in legal documents. These findings highlight the potential of utilizing BERT-based models for automated legal information extraction to support the development of intelligent legal systems and the independence of judiciary in Indonesia.

1. INTRODUCTION

Court decision documents represent official statements issued by judges based on legal reasoning, evidence, and arguments presented during trials. The volume of such documents continues to grow with the increasing number of processed cases. According to data from the Supreme Court of the Republic of Indonesia (www.mahkamahagung.go.id), the average number of new court cases across all levels of jurisdiction reached approximately 893,546 annually between 2020 and 2024.

These documents are typically composed in long, unstructured textual formats characterized by diverse legal terminology, complex syntactic structures, and inconsistent organizational patterns. Such characteristics pose significant challenges for both manual and automated information processing. These conditions make information extraction (IE) particularly difficult, especially when identifying and extracting relevant named entities from court decisions.

IE is vital in identifying and classifying key entities, relationships, and events embedded within unstructured text [1-3]. In legal contexts, IE is essential for accelerating document review processes and enhancing the accuracy of legal decision-making [4, 5]. Automated entity recognition enables legal professionals to efficiently locate relevant information without the need to scrutinize entire documents manually [6] while also reducing workload and enriching legal data analysis through the application of Natural Language Processing (NLP) techniques [7].

Within IE, Named Entity Recognition (NER) is one of the most prominent and widely used tasks. Traditional machine learning approaches, such as Support Vector Machines (SVM) and Conditional Random Fields (CRF), have shown effectiveness in certain domains but rely heavily on handcrafted features, which are time-consuming and lack adaptability [8, 9]. In contrast, deep learning methods like Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Encoder Representations from Transformers (BERT) offer advanced capabilities in capturing contextual dependencies and learning linguistic patterns directly from the data [10, 11]. These attributes make them increasingly suitable for processing complex and nuanced legal language.

Despite existing studies employing SVM, CRF, or Bi-LSTM for general IE tasks, specific applications in the

Indonesian legal domain remain limited [12]. Most prior work has focused on English or other widely used languages, overlooking the unique linguistic and structural challenges of processing Indonesian legal texts. This gap has resulted in limited insights regarding the applicability and optimization of NER models for Indonesian legal documents. Furthermore, although BERT has demonstrated outstanding performance across a wide range of NLP tasks, its effectiveness in highly contextual and terminology-rich domains such as legal documents remain underexplored, particularly in the Indonesian context.

This study addresses these gaps by comparing four NER models, SVM, CRF, Bi-LSTM, and BERT for entity extraction from Indonesian court decision documents. The primary objective is to assess the ability of each model to accurately identify essential legal entities such as names, locations, and domain-specific terms in complex legal texts. By adopting a systematic comparative approach, this research not only provides empirical evidence on model performance but also delivers novel insights into how these methods can be adapted and optimized for legal NLP applications in Indonesia.

By highlighting the application of state-of-the-art models such as BERT and contrasting them with both traditional and deep learning-based approaches, this study aims to contribute significantly to the development of intelligent legal information systems. The findings are expected to support more effective, efficient, and data-driven decision-making processes within the Indonesian judicial system.

2. LITERATURE REVIEW

One of the concepts that can be applied to extracting meaningful information from within a document is information extraction. Information extraction is part of Natural Language Processing (NLP) that can extract essential facts from documents about predetermined types of events, entities, or relationships to build a more meaningful and rich semantic content representation, which can be used to fill in providing structured input in more complex pattern mining [2, 3].

2.1 Learning-based named entity recognition

In the application of information extraction, especially in NER tasks, various methods have been applied, including traditional machine learning-based approaches such as SVM and CRF, as well as modern deep learning-based approaches such as Bi-LSTM and BERT [12].

SVM is one of the popular machine-learning algorithms in NER. This model works by separating data into different classes using hyperplanes in multidimensional space. Although effective in many cases, SVM has limitations regarding complex feature representation and its inability to capture the context of words in sentences [13, 14]. CRF, conversely, is a more sophisticated model that considers dependencies between entities in a sequential context. By leveraging the sequential structure, CRF can handle context better than SVM, making it a good choice for NER tasks where relationships between entities are critical [15-17].

In recent years, deep learning approaches have gained significant attention due to their ability to capture complex patterns in data. Bi-LSTM is an extension of LSTM designed to capture context from both directions, namely from the previous and following words [12]. This condition is critical in NER, where word context often determines entity categorization. BERT is a more advanced model, utilizing the transformer architecture to deeply understand word context by considering all words in a sentence [11]. BERT has shown excellent performance in various NLP tasks, including NER, due to its ability to learn from large and complex data.

2.2 Related work

Research on IE and NER in the legal domain has been explored across multiple jurisdictions, employing methodologies ranging from rule-based systems to machine learning and deep learning approaches. For instance, Jackson et al. [18] developed a legal IE and search system known as History Assistant, which exemplifies the use of IE to generate linkages between new cases and precedent ones by extracting relevant information from court decision documents. The research illustrates the potential of IE to enhance legal research by uncovering implicit case relationships.

Walter and Pinkal [19] conducted studies using computational linguistic techniques to extract definitions and ontological structures from legal corpora. Their work, particularly the Automatic Extraction of Definitions from German Court Decisions, utilizes rule-based approaches to parse and analyze over 6000 legal texts. These approaches are practical in environments where legal syntax and patterns are rigid and well-understood, such as German court decisions. The reliance on manually encoded linguistic rules, however, often limits the scalability and adaptability of such systems to new or more variable data domains.

While rule-based methods have the advantage of transparency and high precision in controlled environments, they require intensive domain expertise to develop and maintain. Moreover, they struggle with the linguistic variability, evolving terminology, and ambiguous structures frequently found in legal texts. In contrast, machine learning and deep learning approaches such as CRF, SVM, Bi-LSTM, and BERT offer the benefit of learning patterns automatically from data, making them more robust to variation and capable of generalizing across broader datasets. However, these methods often require large annotated corpora and careful fine-tuning to perform well in domain-specific tasks such as legal NER.

The integration of domain-specific knowledge into AI/ML systems is critical to ensuring that the resulting solutions are not only effective but also contextually relevant and ethically responsible [20]. This study focuses on the legal domain, where the complexity and sensitivity of language and interpretation demand a nuanced understanding of the field. In this study, incorporating legal domain knowledge into the AI/ML framework is pivotal in enhancing the system's ability to process legal texts accurately and responsibly, aligning technological advancement with domain-specific requirements.

In the Indonesian context, IE research specific to legal documents remains limited. Some initial efforts include the development of an information extraction framework from Indonesian legislation [21], which primarily focused on the rule-based extraction of entities from statutory documents. Additionally, rule-based IE has been applied to general criminal court decision documents [22], although such approaches often face limitations due to the complexity and

inconsistent structure of Indonesian court texts. More recently, deep learning techniques have begun to be explored, such as the use of neural networks for extracting information from criminal court documents [23].

Indonesian legal documents present unique challenges that are not commonly found in legal corpora from other jurisdictions. These challenges include a lack of standardization in document formatting, high variability in vocabulary usage, complex sentence structures, and the frequent embedding of implicit legal reasoning without consistent terminology. Such factors make rule-based and traditional machine learning methods less effective unless extensively customized.

Furthermore, unlike legal NLP in English or German, which benefits from established resources and annotated datasets, the Indonesian legal domain lacks comprehensive corpora, limiting the applicability of off-the-shelf models. This situation creates a research gap and underscores the need for comparative studies evaluating modern approaches such as BERT, particularly their ability to handle the semantic richness and contextual dependencies of the Indonesian legal language.

Thus, although prior studies have explored rule-based and machine-learning methods for legal IE in various languages and jurisdictions, the Indonesian legal domain remains underexplored. There is a critical need to examine how state-of-the-art models like BERT compare with traditional approaches when applied to Indonesian court decisions. This study addresses that gap by conducting a systematic comparative analysis of four NER models SVM, CRF, Bi-LSTM, and BERT, on Indonesian court decision documents, aiming to identify the most effective strategy for extracting key legal entities in this context.

3. METHODOLOGY

The process carried out in this study includes six main stages, namely document provision, preprocessing, annotation, feature, method, and evaluation. The complete research process can be seen in Figure 1.

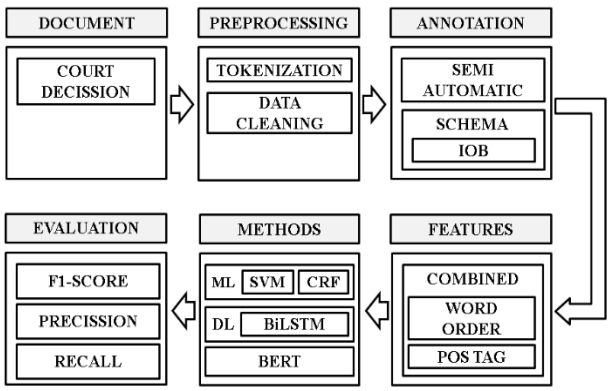

Figure 1. Research process

The first step in this study is to identify the primary data source, namely court decision documents in Indonesia, that can be accessed and taken from the Supreme Court website. These documents contain important legal information such as the defendant's name, date, location, and case number. Legal documents often have non-standard text structures with complex language, rich in technical terms, and varied. Therefore, these documents are a challenge and a valuable source of data to be explored in the NER task. Examples of decision documents and several entities to be extracted can be seen as below:

P U T U S A N

NOMOR: 101/PID.B/2015/PNJAK.TIM.

“DEMI KEADILAN BERDASARKAN KETUHANAN YANG MAHA ESA”

Pengadilan Negeri Jakarta Timur yang memeriksa dan mengadili perkara - perkara pidana dengan acara pemeriksaan biasa dalam peradilan tingkat pertama telah menjatuhkan putusan sebagai berikut atas nama terdakwa:

Nama lengkap: EKO HARTANTO alias JAWIR;
Tempat lahir: Jakarta;

. . . dst

Setelah mendengar pembacaan tuntutan Jaksa Penuntut Umum tertanggal 23 Februari 2015 yang pada pokoknya menuntut terdakwa sebagai berikut:

1 Menyatakan terdakwa EKO HARTANTO alias JAWIR telah terbukti bersalah melakukan tindak pidana sebagaimana diatur dan diancam pidana dalam Pasal 365 ayaat (2) ke - 2 KUHP;

2 Menjatuhkan pidana terhadap terdakwa EKO HARTANTO alias JAWIR berupa pidana penjara selama 2 (dua) tahun dikurangi selama terdakwa berada dalam tahanan sementara;

. . . dst

Demikian diputuskan dalam rapat permusyawaratan Majelis Hakim pada hari : SENIN, tanggal 23 PEBRUARI 2015 , oleh SATRIYO BUDIYONO, SH, M.Hum., selaku Hakim Ketua, BONTOR AROEAN,SH.,MH dan DWI PURWADI, SH. MH., masing-masing selaku Hakim Anggota, dan putusan tersebut diucapkan dalam sidang yang terbuka untuk umum pada hari itu juga : SENIN, tanggal 23 PEBRUARI 2015, oleh Hakim Ketua Majelis tersebut dengan didampingi oleh Hakim-Hakim Anggota, dan dibantu oleh LELY SUCIATI,SH Panitera Pengganti, dengan dihadiri oleh ASRY R. PURWANINGSIH,SH Jaksa Penuntut Umum pada Kejaksaan Negeri Jakarta Timur serta dihadiri oleh Terdakwa;

. . . dst

After the documents are obtained, the next step is preprocessing to ensure the data is ready to use. This process begins with tokenization, dividing the text into small units such as words or phrases, making it easier for further analysis. In addition, data cleaning is performed to remove irrelevant elements, such as special symbols, foreign characters, or unnecessary formatting elements. Normalization is also applied to equalize spelling and terms, thereby reducing ambiguity that may arise from differences in writing.

The annotation stage aims to label relevant entities in the

document. This study uses a semi-automatic approach, where an automated annotation tool helps provide initial labels that are then re-checked and refined by legal experts. The Inside-Outside-Beginning (IOB) scheme is used in the annotation to mark the position of tokens in an entity. The token that is at the beginning of the entity is labeled "B," the token that is inside the entity is labeled "I," and the token that is not included in the entity is labeled "O."

The features used to train the machine learning model are word order and part-of-speech (POS) tagging. Word order helps capture common word order patterns in legal documents. At the same time, POS tagging provides grammatical information such as nouns, verbs, or adjectives relevant to understanding the linguistic context. Combining these two features creates a rich and informative data representation, especially for machine learning models such as SVM and CRF.

This study uses a mixed approach by comparing machine-learning and deep-learning models. SVM is used for simple but effective margin-based classification, while CRF is used for sequential data labeling tasks that require understanding patterns between tokens. The deep learning models used are Bi-LSTM, which captures context from both directions in a sentence, and BERT, which utilizes a transformer architecture to understand word context deeply. This combination of methods allows for a comprehensive evaluation of the effectiveness of each model in the NER task.

The final step is to evaluate the model's performance using standard metrics in NER, namely precision, recall, and F1-score. Precision measures the accuracy of the model's predictions, recall assesses how many entities are successfully recognized, and the F1-score provides a harmonic mean between the two. Evaluations are carried out for each type of entity, such as names, dates, or locations, to identify the strengths and weaknesses of each model in handling complex legal language.

A comparative experiment was conducted to test the superiority of each model. Each model was evaluated using the same dataset, and the results were compared based on predetermined evaluation metrics. This approach allowed the researchers to identify the most effective model in the context of NER in Indonesian court decision documents. In this way, this study not only provides insight into the performance of each model but also provides recommendations for better use of the model in legal applications.

4. RESULT AND DISCUSSION

4.1 Dataset

The total criminal decision data was taken randomly from the decision directory of the Supreme Court website, totaling 3,654 decisions. During the decision-making process, there were some decisions whose documents were incomplete, so 1,000 court decision document data were selected to be processed in the next stage. From these 1,000 data, 200 were taken and annotated by two annotators. So, the dataset used in this study is 200 annotated data.

This study considers several factors in the utilization of 200 annotated datasets. First, the inherent characteristics of court decision documents, which typically span an average of 20 pages per document, contribute to the complexity of dataset processing and annotation. The substantial length and structural variability of these documents present significant challenges in terms of preprocessing and manual labeling efforts. Furthermore, a systematic literature review by Solihin et al. [12] reveals that prior research focused on information extraction (IE) from court decisions most frequently employed datasets ranging between 101 and 500 decision documents. This observation reinforces the relevance and appropriateness of the dataset size used in this study, aligning it with established practices in the legal IE research domain.

From the 200 decisions in the dataset, it was recorded that there were 24,515 sentences and 1,048,576-word tokens, with 25,912 unique words. This dataset is annotated using 24 special entities with tags B_ and I_ + 1 general entity with tag O. A summary of the statistical calculations of the dataset can be seen in Table 1. In contrast, the 24 unique entities used in the court decision documents and their descriptions can be seen in Table 2.

Table 1. Dataset statistics

No.	Description	Count
1	Courts Doc	200
2	Sentences	24,515
3	Word Token	1,048,576
4	Uniq Word	25,912
5	Tag	25

Table 2. List of entities and IOB tags

No.	Entities	IOB Tag	Count
1	Verdict Number	B_VERN	640
		I_VERN	6,607
2	Defendant	B_DEFN	13,034
		I_DEFN	25,283
3	Criminal Action	B_CRIA	102
		I_CRIA	910
4	Article Violation	B_ARTV	1,912
		I_ARTV	16,441
5	Penalties	B_PENA	106
		I_PENA	622
6	Punishment	B_PUNI	178
		I_PUNI	867
7	Date of the Verdict	B_TIMV	135
		I_TIMV	340
8	Presiding Judge	B_JUDP	127
		I_JUDP	865
9	Judge	B_JUDG	315
		I_JUDG	2,145
10	Registrar	B_REGI	138
		I_REGI	767
11	Prosecutor	B_PROS	116
		I_PROS	617
12	Advocate	B_ADVO	322
		I_ADVO	1,010
13	O	O	974,976

The statistical calculations of the court decision dataset used in this study can be seen in Table 2. The most significant number of entities tags other than the O tag is the I_DEFN tag of 25,283, while the smallest number of entity tag occurrences is B_CRIA of 102. From this table, it can also be seen that the number of occurrences of entities or classes is unbalanced, so two measurements can be used, namely Macro Average and Weighted Average, on precision, recall, and F1-score. If you consider entities or classes equivalent, then Macro Average is more appropriate. At the same time, if entities or classes are deemed to have different contributions based on the weight of

occurrence, then it is more suitable to use Weighted Average or Accuracy.

4.2 Experimentation

The most frequently applied methods for IE in the legal domain include rule-based approaches, CRF, SVM, Bi-LSTM and its variants, as well as BERT-based models [12]. Based on this observation, the present study employs four distinct methods, SVM, CRF, Bi-LSTM, and BERT, for comparative evaluation. Specifically, for SVM and CRF, five different feature combinations are applied to enhance performance. In contrast, for BERT, two different pre-trained models are utilized to assess its effectiveness across different linguistic representations.

This study adopts the Macro Average metric rather than the Weighted Average when evaluating precision, recall, and F1-score. Macro Average is considered more appropriate in this context, assuming all entity classes are equally important, providing an unbiased measure across classes. In contrast, a Weighted Average would be more suitable in scenarios where different entity types contribute unequally to the overall performance based on their frequency or importance in the dataset as shown in Table 3.

Table 3. Performance comparison of learning-based models in legal IE

Model	Feature	Precision	Recall	F1-Score
SVM	Without POS	0.11	0.26	0.10
	With POS	0.09	0.19	0.09
	With POS+NEXT	0.09	0.41	0.09
	With POS+PREV	0.13	0.37	0.13
	With POS+PREV+NEXT	0.17	0.41	0.19
CRF	Without POS	0.75	0.50	0.56
	With POS	0.74	0.50	0.56
	With POS+NEXT	0.79	0.56	0.63
	With POS+PREV	0.76	0.55	0.61
	With POS+PREV+NEXT	0.80	0.59	0.65
BiLSTM		0.40	0.44	0.37
BERT	IndoBERT-base-uncased	0.95	0.84	0.93
	Bert-base-Indonesian-522M	**0.96**	**0.85**	**0.94**

SVM exhibits the weakest performance across all configurations, with F1-scores ranging from 0.09 to 0.19. Despite incorporating contextual features such as POS tags and previous/next tokens, SVM cannot capture the linguistic complexity inherent in Indonesian legal documents. This limitation arises primarily due to the model's dependency on static, manually-engineered features and its lack of capacity to model sequential dependencies a crucial aspect in Named Entity Recognition tasks, particularly when dealing with lengthy and syntactically irregular legal texts.

Similarly, the BiLSTM model, although a deep learning architecture, yields suboptimal results (F1-score: 0.37). This may be attributed to several factors: (i) limited training data size, which hampers BiLSTM's ability to generalize; (ii) inability to leverage contextual pretraining as in transformer-based models; and (iii) absence of language-specific adaptations, which are often required in morphologically-rich, domain-specific corpora like legal Indonesian.

CRF shows consistently strong results, especially when enriched with POS and context-based features. The best CRF configuration (POS + PREV + NEXT) achieves an F1-score of 0.65. Its performance demonstrates that statistical sequence labeling models can still be effective when properly tuned and when syntactic cues are embedded. However, CRF still relies heavily on manual feature engineering and lacks the capacity for deep semantic understanding compared to transformer-based models.

The BERT-based models demonstrate the highest performance across all evaluation metrics, with IndoBERT achieving an F1-score of 0.93 and multilingual BERT-base slightly outperforming it with an F1-score of 0.94. Several factors contribute to this performance:

Contextualized Embeddings: Unlike SVM and CRF, BERT generates context-aware token representations, enabling it to distinguish between semantically similar terms based on their usage in different contexts critical for legal texts where polysemy and legal jargon are prevalent.

Pretraining on Large Corpora: Both BERT variants are pre-trained on massive text corpora. IndoBERT, trained explicitly on Indonesian text, brings additional language-specific advantages, while BERT-base benefits from multilingual transferability and robust generalization.

Another study, which also conducted legal entity extraction, recorded the best precision results of 0.80, recall of 0.86, and F1-score of 0.83 [23]. The Bi_LSTM+CRF model obtained this highest value. Compared with the results of this study, it shows that BERT's performance in this study shows superiority in terms of precision and F1-score.

Although both BERT models perform exceptionally well, BERT-base slightly surpasses IndoBERT in all metrics. This subtle difference may be attributed to BERT-base's exposure to broader multilingual contexts, potentially enriching its semantic representation capabilities across languages and domains. Nonetheless, IndoBERT's performance validates the importance of language-specific pretraining, particularly for morphologically and syntactically complex languages such as Indonesian.

These findings highlight the importance of selecting models that are context-aware and language-adaptive for legal IE tasks. The superior performance of transformer-based models in this study emphasizes the shift from traditional and feature-based models toward pre-trained deep learning architectures. This evolution has practical implications for developing intelligent legal information systems capable of extracting key entities with high precision, ultimately supporting automation in judicial workflows.

5. CONCLUSION

This study systematically examined and compared the effectiveness of multiple information extraction models, including SVM, CRF, BiLSTM, and BERT-based architectures, on the named entity recognition task in Indonesian court decision documents. The results demonstrate that transformer-based models, particularly BERT and IndoBERT, significantly outperform traditional machine learning and even neural sequence models, achieving F1-

scores above 0.93. In contrast, feature-based approaches such as SVM and CRF showed moderate to poor performance, emphasizing the limitations of relying on manually engineered features when handling syntactically complex and semantically dense legal texts.

The superior performance of BERT-based models underscores their ability to capture nuanced legal language, adapt to context, and generalize well across various types of named entities. This capability is critical in the legal domain, where accurate entity recognition directly influences downstream tasks such as legal information retrieval, legal summarization, citation network analysis, and argument mining.

From a broader perspective, this research provides empirical evidence supporting the adoption of pre-trained language models for legal information systems in low-resource languages like Indonesian. It is a foundation for future initiatives to modernize legal data processing and judicial transparency through AI. The ability to automatically extract structured legal knowledge from unstructured court rulings holds substantial potential for improving legal analytics, enabling policy research, and assisting legal professionals in navigating large volumes of case law efficiently.

Future work may involve incorporating additional legal-specific pretraining, experimenting with domain adaptation techniques, or exploring hybrid models that combine rule-based logic with deep learning for better interpretability and legal compliance. Furthermore, expanding the annotated dataset to cover a more diverse set of legal domains and jurisdictions within Indonesia would contribute to building more robust and generalizable IE systems.

ACKNOWLEDGMENT

This research was supported and funded by the fundamental research grant of the Indonesian Ministry of Higher Education, Science, and Technology / Kementerian Pendidikan Tinggi, Sains, dan Teknologi (Grant No.: 101/E5/PG.02.00.PL/2024).

REFERENCES

[1] Turmo, J., Ageno, A., Catala, N. (2006). Adaptive information extraction. ACM Computing Surveys (CSUR), 38(2): 4-es. https://doi.org/10.1145/1132956.1132957

[2] Grishman, R. (1997). Information extraction: Techniques and challenges. In Information Extraction a Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, pp. 10-27. https://doi.org/10.1007/3-540-63438-X_2

[3] Riloff, E. (1999). Information extraction as a stepping stone toward story understanding. Understanding language understanding: Computational models of reading, 435-460. https://doi.org/10.7551/mitpress/6981.003.0015

[4] Hwang, W., Eom, S., Lee, H., Park, H.J., Seo, M. (2022). Data-efficient end-to-end information extraction for statistical legal analysis. arXiv Preprint arXiv:2211.01692. https://doi.org/10.18653/v1/2022.nllp-1.12

[5] Mistica, M., Zhang, G.Z., Chia, H., Shrestha, K.M., Gupta, R.K., Khandelwal, S., Paterson, J., Baldwin, T., Beck, D. (2020). Information extraction from legal documents: A study in the context of common law court judgements. In Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association, pp. 98-103. https://aclanthology.org/2020.alta-1.12

[6] Zadgaonkar, A.V., Agrawal, A.J. (2021). An overview of information extraction techniques for legal document analysis and processing. International Journal of Electrical & Computer Engineering, 11(6): 5450-5457. https://doi.org/10.11591/ijece.v11i6.pp5450-5457

[7] Remus, D., Levy, F. (2017). Can robots be lawyers: Computers, lawyers, and the practice of law. SSRN. https://doi.org/10.2139/ssrn.2701092

[8] Califf, M.E. (1999). Relational learning of pattern-match rules for. In Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp. 328. https://aclanthology.org/W97-1002.

[9] Abdallah, Z.S., Carman, M., Haffari, G. (2017). Multi-domain evaluation framework for named entity recognition tools. Computer Speech & Language, 43: 34-55. https://doi.org/10.1016/j.csl.2016.10.003

[10] Chen, J., Huang, Y., Yang, F., Li, C. (2020). A novel named entity recognition approach of judicial case texts based on BiLSTM-CRF. In 2020 12th International Conference on Advanced Computational Intelligence (ICACI), Dali, China, pp. 263-268. https://doi.org/10.1109/ICACI49185.2020.9177731

[11] Kenton, M.C., Kristina, L., Devlin, J. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[12] Solihin, F., Budi, I., Aji, R.F., Makarim, E. (2021). Advancement of information extraction use in legal documents. International Review of Law, Computers & Technology, 35(3): 322-351. https://doi.org/10.1080/13600869.2021.1964225

[13] Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press.

[14] Vapnik, V.N. (2013). The Nature of Statistical Learning Theory. Springer New York, NY. https://doi.org/10.1007/978-1-4757-3264-1

[15] Sutton, C., McCallum, A. (2012). An introduction to conditional random fields. Foundations and Trends® in Machine Learning, 4(4): 267-373. https://doi.org/10.1561/2200000013

[16] Sarawagi, S., Cohen, W.W. (2004). Semi-markov conditional random fields for information extraction. NIPS'04: Proceedings of the 18th International Conference on Neural Information Processing Systems, Cambridge, MA, United States.

[17] Liu, K., El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. Automation in Construction, 81: 313-327. https://doi.org/10.1016/j.autcon.2017.02.003

[18] Jackson, P., Al-Kofahi, K., Tyrrell, A., Vachher, A. (2003). Information extraction from case law and retrieval of prior cases. Artificial Intelligence, 150(1-2): 239-290. https://doi.org/10.1016/S0004-3702(03)00106-1

[19] Walter, S., Pinkal, M. (2006). Automatic extraction of definitions from German court decisions. In Proceedings of the Workshop on Information Extraction Beyond the Document, pp. 20-28. http://doi.org/10.3115/1641408.1641411
[20] Miller, T., Durlik, I., Łobodzińska, A., Dorobczyński, L., Jasionowski, R. (2024). AI in context: Harnessing domain knowledge for smarter machine learning. Applied Sciences, 14(24): 11612. https://doi.org/10.3390/app142411612
[21] Hartadi, B., Budi, I. (2017). Punishment provision extraction from Indonesian law texts with knowledge acquisition rules. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, Indonesia, pp. 204-209. https://doi.org/10.1109/ICACSIS.2017.8355034
[22] Solihin, F., Budi, I. (2018). Recording of law enforcement based on court decision document using rule-based information extraction. In 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Yogyakarta, Indonesia, pp. 349-354. https://doi.org/10.1109/ICACSIS.2018.8618187
[23] Nuranti, E.Q., Yulianti, E. (2020). Legal entity recognition in Indonesian court decision documents using Bi-LSTM and CRF approaches. In 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, pp. 429-434. http://doi.org/10.1109/ICACSIS51025.2020.9263157

NOMENCLATURE

BERT	Bidirectional Encoder Representation with Transformers
Bi-LSTM	Bi-directional Long Short-Term Memory
CRF	Conditional Random Forest
IE	Information Extraction
NLP	Natural Language Processing
SVM	Support Vector Machine