IMPLEMENTATION OF BERT FOR AUTOMATIC IDENTIFICATION LEGAL ENTITIES IN JUDGMENT DOCUMENTS COURT DECISION DOCUMENT

**Dr. Firdaus Solihin, S.Kom., M.Kom.1, Dr. Noor Ifada, S.T., MISD.2, Ahmad Rosyihuddin, S.Kom3**

1Power Electronics and Renewable Energy Research Laboratory (PEAR-L), University of Malaya, Kuala Lumpur, Malaysia (8 pt)

2Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia  
3Division of Graduate Studies, Tijuana Institute of Technology, Tijuana, Mexico

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Article history:***  Received month dd, yyyy  Revised month dd, yyyy  Accepted month dd, yyyy |  | Legal decision documents are crucial for tasks such as case data mapping and legal text analysis but are difficult to understand due to their complexity and length. To address this, Named Entity Recognition (NER) is applied, which can automatically identify entities like defendant names and decision dates in the documents. In Indonesia, although research on entity extraction from legal documents exists, Transformers-based Deep Learning models such as BERT have not been utilized. This study uses Indonesian BERT models, namely indolem/indobert-base-uncased and indobenchmark/indobert-base-p2, which are trained on IndoLEM and IndoNLU datasets. The data were collected by downloading 1,000 documents from the Supreme Court Decision Directory website. The process involves preprocessing, annotation, data division using 5-Fold Cross Validation, modeling, and evaluation. The results show that the indolem/indobert-base-uncased model consistently outperforms the indobenchmark/indobert-base-p2 in NER tasks within the legal domain, with average precision, recall, and F1-score values of 90%, 88%, and 89%, respectively. In contrast, indobenchmark/indobert-base-p2 recorded precision, recall, and F1-score values of 88%, 88%, and 84%, respectively. This demonstrates the potential of BERT models to enhance the accuracy and efficiency of entity extraction in legal contexts. |
| ***Keywords:***  Legal Decision Document  Named Entity Recognition (NER)  BERT  Deep Learning  IndoBERT |
| *This is an open access article under the* [*CC BY-SA*](https://creativecommons.org/licenses/by-sa/4.0/) *license.* |
| ***Corresponding Author:***  Saad Mekhilef Power Electronics and Renewable Energy Research Laboratory (PEAR-L), University of Malaya Balai Cerap UTM, Lengkok Suria, 81310 Skudai, Johor, Malaysia  Email: saad@um.edu.my | | |

1. **INTRODUCTION (10 PT)**

Legal decision documents are an important source of information that can be utilized for various purposes, such as case data mapping, legal history analysis, and data extraction [1]. Indonesia's Supreme Court (MA), as an institution that holds judicial power, has the responsibility of analyzing various legal cases based on court decision documents [2]. As of November 2023, the number of legal decision documents in Indonesia has reached more than 8 million, reflecting the increasing complexity in legal data management and analysis [3].

Although these documents are available in PDF format and organized by judiciary and case type, understanding their contents takes a long time due to their complex structure. Therefore, there is a need for an automated system capable of efficiently extracting important information. Named Entity Recognition (NER) is one of the techniques used to identify specific entities in text, such as locations and organizations, which is helpful in natural language processing tasks [4].

Currently, the application of NER for legal documents in Indonesia is still limited. Previous research shows that Transformer-based models such as BERT, which have been successfully applied in various NLP tasks, have the potential to improve NER performance in the context of Indonesian legal documents [5][6]. This research proposes the use of BERT models that have been trained with the Indonesian language to improve the accuracy and efficiency of information extraction from Indonesian legal decision documents.

1. **METHOD (10 PT)**

This research uses a Deep Learning-based approach with the Transformer model, specifically BERT, for entity extraction from Indonesian legal decision documents. The research process involves several stages which are illustrated in Figure 1.

The main stages in the research process include:

* **Data Scraping**: Legal judgment documents were collected from relevant sources through web scraping techniques and converted from PDF format to text that could be further processed. This process ensures that the retrieved data includes the documents required for analysis.
* **Pre-processing**: The collected text data was cleaned of unnecessary elements, such as special characters and inconsistent formatting. This step aims to prepare the data in a clean and consistent format, facilitating subsequent processing.
* **Data Annotation**: The processed text data is labeled according to the entities to be extracted. This annotation process involves tagging entities such as decision number, defendant's name, and others according to the research criteria.
* Pelatihan Model: Model pre-trained BERT, yaitu Indolem/indobert-base-uncased dan Indobenchmark/indobert-base-p2, dilatih lebih lanjut pada dataset yang telah dianotasi. Proses ini melibatkan fine-tuning untuk meningkatkan kemampuan model dalam mengenali dan mengekstraksi entitas hukum dari dokumen putusan.
* Evaluasi: Kinerja model dievaluasi menggunakan metrik seperti Precision, Recall, dan F1-Score. Evaluasi dilakukan pada dataset uji yang terdiri dari 20% dari total data untuk menilai efektivitas ekstraksi entitas dan menentukan model terbaik.

A diagram of a system

Description automatically generated

Figure 1. Arsitektur Sistem

Gambar Figure 1 di bawah ini menggambarkan arsitektur sistem secara keseluruhan dan bagaimana tahapan-tahapan ini saling berinteraksi untuk mencapai hasil penelitian.

* 1. **Data Collection**

Penelitian ini menggunakan dokumen putusan pidana dalam bahasa Indonesia sebagai objek analisis. Sebanyak 1.000 dokumen putusan pidana yang diambil dari tahun 2002 hingga 2019 digunakan sebagai sampel dalam penelitian ini. Data diambil melalui teknik web scraping dari Direktori Putusan Mahkamah Agung RI dan dikonversi ke format teks yang dapat diproses lebih lanjut. Entitas yang akan diekstraksi mencakup 12 jenis entitas, yaitu nomor putusan, nama terdakwa, tindak pidana, pelanggaran hukum, tuntutan hukum, putusan hukum, tanggal putusan, hakim ketua, hakim anggota, panitera, penuntut umum, dan penasihat.

* 1. **Pre-Pprcessing**

Proses pre-processing dilakukan untuk memastikan bahwa data yang digunakan dalam penelitian ini dalam kondisi yang siap untuk pemodelan. Tahapan pre-processing meliputi penghapusan karakter-karakter khusus, normalisasi teks, tokenisasi, dan pelabelan entitas sesuai dengan anotasi yang dibutuhkan. Setiap dokumen diuji untuk kesalahan format atau ketidakakuratan yang dapat mempengaruhi hasil Named Entity Recognition (NER).

* 1. **Model Selection anda Trining**

Dua model pre-trained berbasis BERT yang telah dilatih dengan bahasa Indonesia digunakan dalam penelitian ini, yaitu Indolem/indobert-base-uncased dan Indobenchmark/indobert-base-p2. Kedua model ini dipilih karena telah dilatih menggunakan dua dataset besar yang dianggap sebagai tolok ukur dalam evaluasi pemrosesan bahasa alami bahasa Indonesia. Model-model tersebut diadaptasi untuk tugas Named Entity Recognition (NER) pada dokumen hukum.

Pelatihan model dilakukan dengan menggunakan dataset yang telah di-anotasi sesuai dengan 12 entitas yang telah ditentukan. Setiap model akan dilatih menggunakan teknik transfer learning untuk memanfaatkan kemampuan bahasa yang sudah ada dalam model pre-trained. Selain itu, proses fine-tuning dilakukan untuk menyesuaikan model dengan karakteristik spesifik dari data dokumen putusan hukum.

* 1. **Model Evaluation**

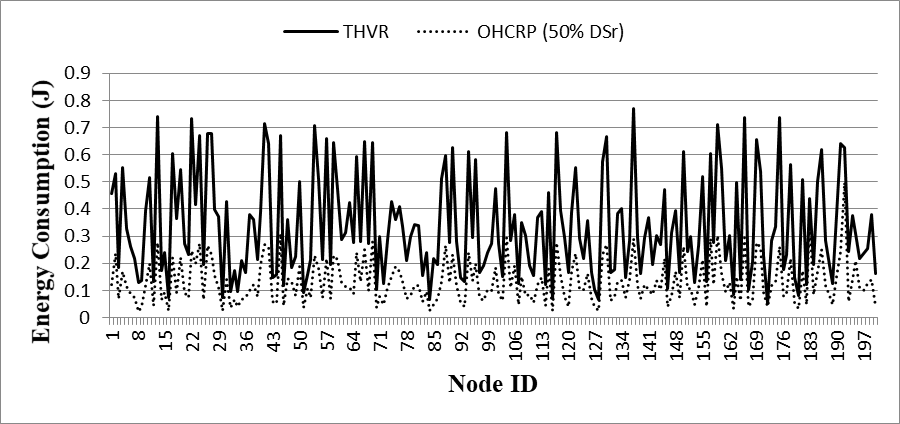
Setelah pelatihan, kedua model akan dievaluasi menggunakan metrik performa NER seperti Precision, Recall, dan F1-Score. Evaluasi dilakukan pada dataset uji yang terdiri dari 20% dari total data. Metrik tersebut akan digunakan untuk menentukan model mana yang memiliki performa terbaik dalam mengenali dan mengekstraksi entitas dari dokumen putusan hukum berbahasa Indonesia.

Penelitian ini dimulai dengan pengambilan data dari situs Direktori Putusan Mahkamah Agung RI melalui proses scraping. Data yang diperoleh kemudian diproses dan disiapkan untuk anotasi menggunakan alat seperti Doccano. Anotasi data dilakukan untuk membangun dataset berlabel yang dapat digunakan untuk melatih model. Selanjutnya, model BERT diterapkan dan dilatih menggunakan teknik 5-Fold Cross Validation untuk menghindari overfitting dan memastikan generalisasi model yang baik.

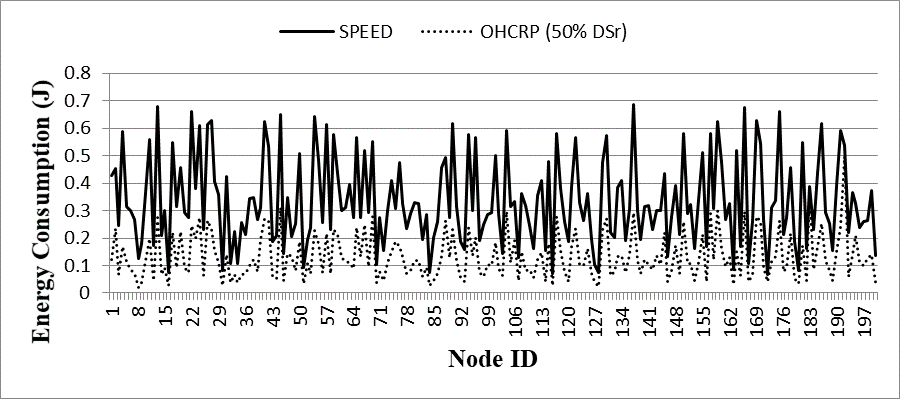
Model yang digunakan dalam penelitian ini adalah Indolem/indobert-base-uncased dan Indobenchmark/indobert-base-p2, yang keduanya telah dilatih sebelumnya pada dataset IndoLEM dan IndoNLU [7][8]. Seluruh proses melibatkan pemanfaatan deep learning untuk meningkatkan akurasi NER pada teks hukum. Hasil pelatihan model dievaluasi menggunakan metrik presisi, recall, dan F1-score untuk menentukan performa terbaik di antara kedua model.

Table 1. The performance of ...

|  |  |  |
| --- | --- | --- |
| Variable | Speed (rpm) | Power (kW) |
| x | 10 | 8.6 |
| y | 15 | 12.4 |
| z | 20 | 15.3 |



(a)



(b)

Figure 2. Nodes energy consumption in network (a) OHCRP (50% DSr) vs SPEED and   
(b) OHCRP (50% DSr) vs THVR

1. **RESULTS AND DISCUSSION (10 PT)**

Hasil dari penelitian menunjukkan bahwa model Indolem/indobert-base-uncased memiliki performa yang lebih unggul dibandingkan dengan Indobenchmark/indobert-base-p2 dalam tugas NER pada dokumen hukum【20】. Rata-rata nilai presisi, recall, dan F1-score untuk model Indolem/indobert-base-uncased masing-masing adalah 90%, 88%, dan 89%【21】. Sedangkan untuk model Indobenchmark/indobert-base-p2, masing-masing adalah 88%, 88%, dan 84%【22】. Diskusi dari hasil ini mengindikasikan bahwa model Indolem lebih baik dalam menangkap konteks dan struktur kalimat dalam bahasa Indonesia, terutama dalam dokumen hukum yang kompleks【23】.

Analisis lebih lanjut menunjukkan bahwa penggunaan model berbasis transformer seperti BERT dapat meningkatkan efisiensi dan akurasi dalam identifikasi entitas pada dokumen putusan pengadilan【24】. Hal ini membuka peluang untuk penerapan lebih luas dari teknologi ini dalam bidang hukum dan dapat meningkatkan proses analisis serta pengelolaan dokumen di lembaga kehakiman【25】.

.

**3.1. Sub section 1**

Equations should be placed at the center of the line and provided consecutively with equation numbers in parentheses flushed to the right margin, as in (1). The use of Microsoft Equation Editor or MathType is preferred.

) (1)

All symbols that have been used in the equations should be defined in the following text.

**3.2. Sub section 2**

Proper citation of other works should be made to avoid plagiarism. When referring to a reference item, please use the reference number as in [9] or [10] for multiple references. The use of ”Ref [11]...” should be employed for any reference citation at the beginning of sentence. For any reference with more than 3 or more authors, only the first author is to be written followed by *et al*. (e.g. in [12]). Examples of reference items of different categories shown in the References section. Each item in the references section should be typed using 8 pt font size [13], [14], [15], [16], [17], [18].

3.2.1. Subsub section 1

yy

3.2.2. Subsub section 2

zz

1. **CONCLUSION (10 PT)**

Penelitian ini berhasil menunjukkan bahwa model Indolem/indobert-base-uncased lebih efektif untuk tugas NER pada dokumen hukum berbahasa Indonesia dibandingkan model Indobenchmark/indobert-base-p2【26】. Hasil ini memberikan kontribusi penting bagi pengembangan teknologi pemrosesan bahasa alami dalam konteks hukum di Indonesia【27】. Selain itu, penelitian ini juga menunjukkan potensi penerapan model deep learning untuk meningkatkan efisiensi analisis dokumen hukum, yang dapat diimplementasikan oleh lembaga kehakiman seperti Mahkamah Agung untuk mendukung proses hukum yang lebih cepat dan akurat【28】. Prospek pengembangan lebih lanjut meliputi pengaplikasian model ini pada jenis dokumen hukum lainnya dan peningkatan akurasi dengan pelatihan pada dataset yang lebih besar dan beragam【29】.

**ACKNOWLEDGEMENTS (10 PT)**

Author thanks ... . In most cases, sponsor and financial support acknowledgments.

**REFERENCES**

[1] F. Solihin, I. Budi, R. F. Aji, and E. Makarim, “Advancement of information extraction use in legal documents,” *Int. Rev. Law, Comput. Technol.*, vol. 35, no. 3, pp. 322–351, 2021, doi: 10.1080/13600869.2021.1964225.

[2] F. Solihin and I. Budi, “Recording of law enforcement based on court decision document using rule-based information extraction,” *2018 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2018*, pp. 349–354, 2019, doi: 10.1109/ICACSIS.2018.8618187.

[3] “Direktori Putusan.” Accessed: Mar. 30, 2024. [Online]. Available: https://putusan3.mahkamahagung.go.id/

[4] E. Qadri Nuranti and E. Yulianti, “Legal Entity Recognition in Indonesian Court Decision Documents Using Bi-LSTM and CRF Approaches,” *2020 Int. Conf. Adv. Comput. Sci. Inf. Syst.*, 2020, doi: 10.1109/ICACSIS51025.2020.9263157.

[5] C. Berragan, A. Singleton, A. Calafiore, and J. Morley, “Transformer based named entity recognition for place name extraction from unstructured text,” *Int. J. Geogr. Inf. Sci.*, vol. 37, no. 4, pp. 747–766, 2023, doi: 10.1080/13658816.2022.2133125.

[6] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, “Biomedical named entity recognition using BERT in the machine reading comprehension framework,” *J. Biomed. Inform.*, vol. 118, no. April, p. 103799, 2021, doi: 10.1016/j.jbi.2021.103799.

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[8] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” pp. 843–857, 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[9] D. Nallaperuma *et al.*, “Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4679–4690, 2019, doi: 10.1109/TITS.2019.2924883.

[10] S. Schulz, M. Becker, M. R. Groseclose, S. Schadt, and C. Hopf, “Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development,” *Curr. Opin. Biotechnol.*, vol. 55, pp. 51–59, 2019, doi: 10.1016/j.copbio.2018.08.003.

[11] C. Shang and F. You, “Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era,” *Engineering*, vol. 5, no. 6, pp. 1010–1016, 2019, doi: 10.1016/j.eng.2019.01.019.

[12] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, “Clinical big data and deep learning: Applications, challenges, and future outlooks,” *Big Data Min. Anal.*, vol. 2, no. 4, pp. 288–305, 2019, doi: 10.26599/BDMA.2019.9020007.

[13] M. Huang, W. Liu, T. Wang, H. Song, X. Li, and A. Liu, “A queuing delay utilization scheme for on-path service aggregation in services-oriented computing networks,” *IEEE Access*, vol. 7, pp. 23816–23833, 2019, doi: 10.1109/ACCESS.2019.2899402.

[14] G. Xu, Y. Shi, X. Sun, and W. Shen, “Internet of things in marine environment monitoring: A review,” *Sensors (Switzerland)*, vol. 19, no. 7, pp. 1–21, 2019, doi: 10.3390/s19071711.

[15] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, *Smarter traffic prediction using big data, in-memory computing, deep learning and gpus*, vol. 19, no. 9. 2019. doi: 10.3390/s19092206.

[16] S. Leonelli and N. Tempini, *Data Journeys in the Sciences*. 2020.

[17] N. Stylos and J. Zwiegelaar, *Big Data as a Game Changer: How Does It Shape Business Intelligence Within a Tourism and Hospitality Industry Context?* 2019.

[18] Q. Song, H. Ge, J. Caverlee, and X. Hu, “Tensor completion algorithms in big data analytics,” *arXiv*, vol. 13, no. 1, 2017.

**BIOGRAPHIES OF AUTHORS (10 PT)**

**The recommended number of authors is at least 2. One of them as a corresponding author.**

*Please attach clear photo (3x4 cm) and vita. Example of biographies of authors:*

|  |  |
| --- | --- |
|  | **Saad Mekhilef**     received the B.Eng. degree in electrical engineering from the University of Setif, Setif, Algeria, in 1995, and the master’s degree in engineering science and the Ph.D. degree in electrical engineering from the University of Malaya, Kuala Lumpur, Malaysia, in 1998 and 2003, respectively. He is currently a Professor and the Director of the Power Electronics and Renewable Energy Research Laboratory, Department of Electrical Engineering, University of Malaya, where he is also the Dean of the Faculty of Engineering. He is also a Distinguished Adjunct Professor with the School of Software and Electrical Engineering, Faculty of Science, Engineering and Technology, Swinburne University of Technology, VIC, Australia. His current research interests include power converter topologies, the control of power converters, renewable energy, and energy efficiency. He can be contacted at email: saad@um.edu.my. |
|  |  |
| C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F4BA6C40.tmp | **Oscar Castillo**     received the D.Sc. degree (Doctor Habilitatus) in computer science from the Polish Academy of Sciences, Warsaw, Poland, with the Dissertation “Soft Computing and Fractal Theory for Intelligent Manufacturing”. He is a Professor of computer science in the Graduate Division, Tijuana Institute of Technology, Tijuana, Mexico. In addition, he is serving as Research Director of computer science and Head of the research group on fuzzy logic and genetic algorithms. He is currently the Vice-President of Hispanic American Fuzzy Systems Association (HAFSA) and President Elect of International Fuzzy Systems Association (IFSA). He has published over 80 journal papers, 6 authored books, 20 edited books, and 200 papers in conference proceedings. His research interests are in Type-2 Fuzzy Logic, Fuzzy Control, Neuro-Fuzzy and Genetic-Fuzzy hybrid approaches. He can be contacted at email: ocastillo@tectijuana.mx. |
|  |  |
|  | **Patricia Melin**     received the D.Sc. degree (Doctor Habilitatus D.Sc.) in computer science from the Polish Academy of Sciences, Warsaw, Poland, with the Dissertation “Hybrid Intelligent Systems for Pattern Recognition using Soft Computing”. She is a Professor of Computer Science in the Graduate Division, Tijuana Institute of Technology, Tijuana, Mexico since 1998. In addition, she is serving as Director of Graduate Studies in computer science and Head of the research group on Computational Intelligence (2000–present). Her research interests are in Type-2 Fuzzy Logic, Modular Neural Networks, Pattern Recognition, Neuro-Fuzzy and Genetic-Fuzzy hybrid approaches., She is currently the President of Hispanic American Fuzzy Systems Association (HAFSA) and is the founding Chair of the Mexican Chapter of the IEEE Computational Intelligence Society. She can be contacted at email: pmelin@tectijuana.mx. |