Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimization of Fine-Tuned DistilBERT Model for Classification of Status And Type of Laws and Regulations

Yorissa Silviana[1], Rianto[2], Vega Purwayoga[3]
[123] *Informatics Study Program, Faculty of Engineering, Siliwangi University*
217006502@student.unsil.ac.id[1], rianto@unsil.ac.id[2], vega.purwayoga@unsil.ac.id[3]

ARTICLE INFO

Received: 05 Mar 2025

Revised: 26 Apr 2025

Accepted: 05 May 2025

ABSTRACT

Data management of laws and regulations is critical to support an efficient legal information system. Still, the complexity of legal language, the diversity of document structures, and the large volume of data are the main challenges in the automatic classification process. This research aims to optimize the DistilBERT model through a fine-tuning approach with a multi-task learning scheme to predict two labels simultaneously, namely regulation status (Applicable / Not Applicable) and type/form of regulation. The research stages include data collection, preprocessing, model training, and model evaluation. The model achieved high performance on the two classification tasks, with 96% accuracy, 94% precision, 96% recall, and 94% f1-score for Regulation Status classification, as well as 100% perfect results on all evaluation metrics for Regulation Type/Shape classification, demonstrating the accuracy and reliability of the model in understanding and classifying legal documents as a whole. This finding confirms that the optimized model is highly reliable in the classification of the status of laws and regulations.

Keywords: Classification, DistilBERT, Fine-tuning, Natural Language Processing, Optimization

INTRODUCTION

Legislation is a very important legal instrument in maintaining social, political, and economic order in a country. The classification of legislation is essential to support quick and accurate decision-making, for the government, legal practitioners, and the general public (Petrakova, 2022). However, the complexity of the text contained in regulatory documents, as well as the diversity of formats and terms used, often complicates the data processing process (Alexopoulos et al., 2024). This is especially challenging when the volume of documents being managed is very large, requiring technology that can automatically classify or extract information from these texts. Advances in artificial intelligence technology, particularly in natural language processing (NLP), offer effective solutions to these challenges. The Bidirectional Encoder Representations from Transformers (BERT) model, which has been proven capable of capturing the contextual meaning of words in depth, has become one of the most widely used models in various NLP applications (Fang et al., 2023). However, with its large model size and high computational requirements, the use of BERT at scale can be inefficient (Dong, 2024). Research on the application of BERT and other transformer variants in the legal domain has been conducted extensively. For example, (Akca et al., 2022) research has developed a model for classifying Turkish legal documents using transformers and domain adaptation. Other (Haque et al., 2022) research has developed pre-trained language models specifically for the Indian legal domain, including a vocabulary tailored to Indian legal texts. In addition (Quevedo et al., 2022), BERT models have been used for answering questions related to legal documents in the PolicyQA system. The use of DistilBERT has also been tested in various domains, including the identification

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

of hateful content in social media (Kumar et al., 2020) and sentiment analysis of public opinion related to COVID-19 (Jojoa et al., 2024). Previous studies (Bambroo & Awasthi, 2021), such as those conducted by , show that using DistilBERT with domain-specific pre-training can improve accuracy in legal document classification. This study proposes the use of a fine-tuning strategy that is more focused on specific domains. However, no one has explored the use of this model in large-scale regulation classification, let alone in the Indonesian legal context. As an alternative, DistilBERT, a variant of BERT with a smaller model size and better efficiency, emerges as an attractive option. DistilBERT retains 97% of BERT's performance but with 40% fewer parameters and 60% faster processing, making it more suitable for applications with computational limitations and a need for fast processing (Sanh et al., 2019). DistillBERT is also capable of transferring semantic knowledge well, with a size 30% smaller and a speed 83% faster than the ELMo model for comparison (Berfu B et al., 2020). Although DistilBERT has proven to be more efficient and faster than BERT, research evaluating its performance in processing regulatory documents across different countries remains limited. The DistilBERT model needs to be further tested in the context of regulatory classification, which is a key focus in Indonesian regulatory documents, such as distinguishing the status and type of regulations. Given the importance of evolving and updated regulations in the legal context, optimizing the DistilBERT model to address this complexity can significantly contribute to improving the efficiency of legal document processing. In addition, the optimization that forms the basis of this research involves the use of regularization techniques such as dropout and adaptive learning rate reduction to prevent overfitting (Liu et al., 2023) (Xie et al., 2022). The proposed optimization provides an overview of the technical steps that can be used to improve model performance.

METHODS

This research is quantitative research with an experimental approach. The experimental approach is used because this study aims to determine the cause-and-effect relationship between the variables involved, such as the dataset, model architecture, and model parameters (Lê & Schmid, 2022). The novelty of the proposed research lies in the use of legal regulation data for fine-tuning using the DistilBERT model, to classify legal regulations, particularly in the context of Indonesian law. The research stages are presented as a whole in **Figure 1**.

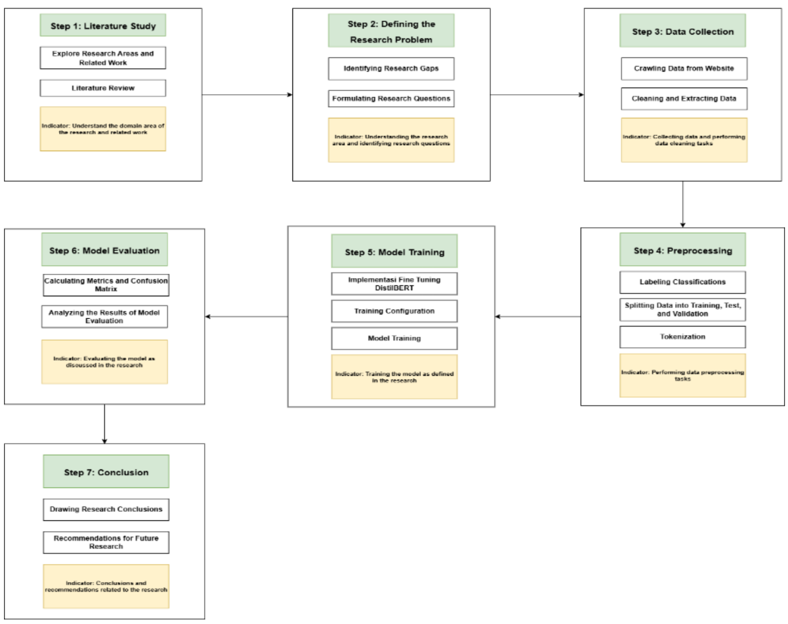

Figure 1. Research Stages

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Literature Study

A literature review is a stage for collecting concepts, theories, and data from various sources related to the research. A literature review is conducted to understand concepts and theories related to the research, such as theories regarding BERT architecture, DistilBERT, and related laws and regulations. Information is gathered using secondary sources (web, journals, e-books, articles, and others). Additionally, during the literature review stage, a "review paper" or analysis of previous research related to the topic of the study is conducted.

Defining the Research Problem

The definition of the research problem is a further step based on the analysis of previous research. Based on the analysis of previous research, gaps in the research are identified, or shortcomings are identified, so that improvements can be made. After identifying the gaps or shortcomings in previous research, research questions are then formulated to determine the objectives of the research to be conducted.

Data Collection

Data collection is the first step before conducting an experiment. The collected dataset will be used to create and test models. The source selected for data collection is the official government website at https://peraturan.go.id, which provides a large amount of legal data.

Preprocessing

Preprocessing is the preparation stage before conducting experiments. In this stage, several text attributes from the dataset are combined into a single text column to form a comprehensive representation of the context of the regulations. The target labels for classification consist of two types: status and abbreviations of regulation types/forms, which are converted into numerical labels. Next, the data is divided into training data and test data with a ratio of 80:20, both for status labels and types. The text data is then tokenized using the tokenizer from the DistilBERT pre-trained model, with a maximum length of 512 tokens and automatic padding.

Model Training

Model training is the next step after preprocessing to implement the DistilBERT model. The implementation uses a multi-task learning approach with a modified DistilBERT model to predict two outputs simultaneously, namely binary classification for status and multi-class classification for the type/form of regulation. At this stage, the model consists of two separate classification layers, each handling a different classification task, and optimized simultaneously by calculating the sum of two CrossEntropy loss functions. The training process was conducted using a customized Huggingface Trainer (MultiTaskTrainer) to calculate the loss from two labels simultaneously and sum them for joint optimization. Training parameters include 3 epochs, a batch size of 16, a learning rate of 5e-5, and the use of dropout to reduce overfitting.

Model Evaluation

Model evaluation is the measurement and collection of results. At this stage, the fine-tuned model will be tested. The evaluation is performed on two separate classification outputs using accuracy, precision, recall, F1-score, and confusion matrix metrics. Prediction results for the test data are extracted using a special prediction function and then compared with the actual labels. In addition, model performance is visualized with training loss and validation loss graphs to observe training stability. A single text prediction function is also provided to test classification results on new inputs in real-time, including predictions of status and type of regulations along with their confidence scores.

Conclusion

Conclusion is the final stage of the research process, which provides an overview of the data analysis and model evaluation results covering the entire study. In addition, conclusions serve as a summary of the main findings of the research, which can be used as a basis for recommendations, decision-making, or further development.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

RESULTS AND DISCUSSION

Base Model

In this research, the model used is DistilBERT which is derived from the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). BERT is trained to be able to perform alternating processing so that the model can be fine-tuned for various needs easily because it only requires one additional output layer (Kurtic et al., 2022). The BERT architecture shown in **Figure 2** shows the advantages of BERT which is simple by using only encoder transformers but with a high GLUE score of 80.5%.

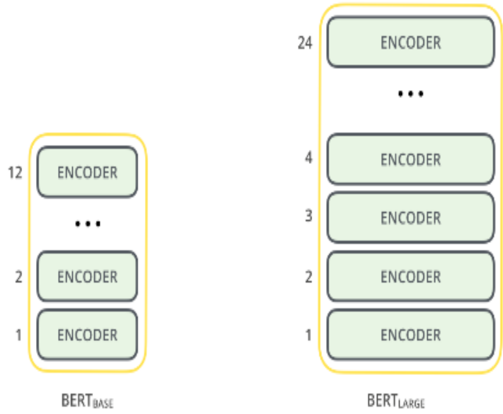

Figure 2. BERT Architecture

BERT requires high resources, so variants that only require small resources are made such as medium, small, mini, and tiny variants (I. Turc, 2019). **Table 1** shows the number of parameters generated by BERT variants.

Table 1. Parameters of BERT Model Variants

Variant	Parameter
BERT-base	110 million
BERT-large	340 million
BERT-medium	41,7 million
BERT-small	29,1 million
BERT-mini	11,3 million
BERT-tiny	4,4 million

This method of reducing resources is called distillation from a larger model (teacher) to a smaller model (student). This method was also applied to produce the DistilBERT model (Lin & Nuha, 2023). This model has 40% fewer parameters than BERT-base and runs 60% faster while still maintaining more than 95% of BERT's performance. DistilBERT is widely used in various needs such as sentiment analysis (Maheshwari et al., 2023)(Azhar & Latif, 2022)(Joshy & Sundar, 2022), question answering (Mollá, 2022)(Kumari et al., 2022), and other tasks. The DistilBERT model architecture can be seen in **Figure 3**.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

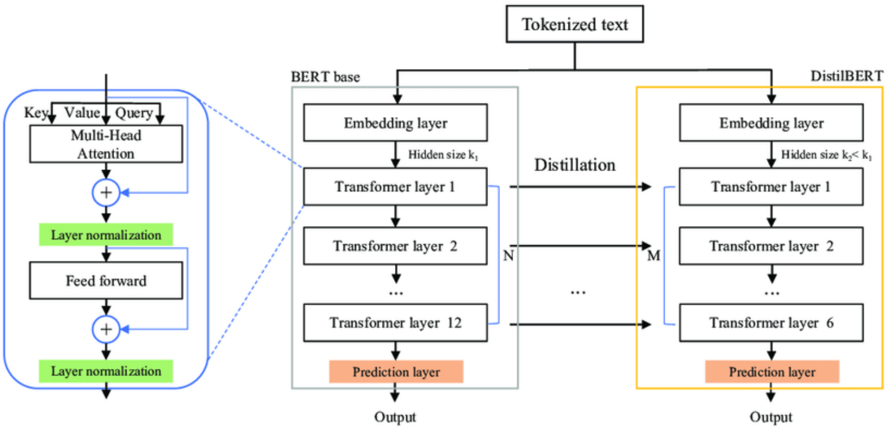

Figure 3. DistilBERT Architecture

The DistilBERT model used in this research can have high accuracy in evaluation but also requires lower resources when fine-tuning the model. This makes DistilBERT an ideal choice for applications that require superior performance while still being economical in memory usage and computation time.

Dataset

To be able to produce a language model for existing laws in Indonesia, data is needed from legal sources in Indonesia. Legal sources in Indonesia are divided into several sources such as the constitution, laws and regulations, courts, and customary law (Firma et al., 2019). In this study, the source of legal data taken comes from stat-utory data, in state regulations the types of regulations that exist in the hierarchy of state law are the Constitution, MPR Decree, Law, Presidential Regulation, UUDRT, Government Regulation, Presidential Decree, Presidential Instruction, Ministerial Regulation, State Agency Regulation and Regional Regulation (Simanjuntak, 2019).

Data will be crawled from specified sources in PDF format. The data will then be extracted and cleaned, after which it will be formatted according to training requirements. The regulations will be re-selected based on their significance to the law so that they can be crawled and extracted according to the flow described in **Figure 4**.

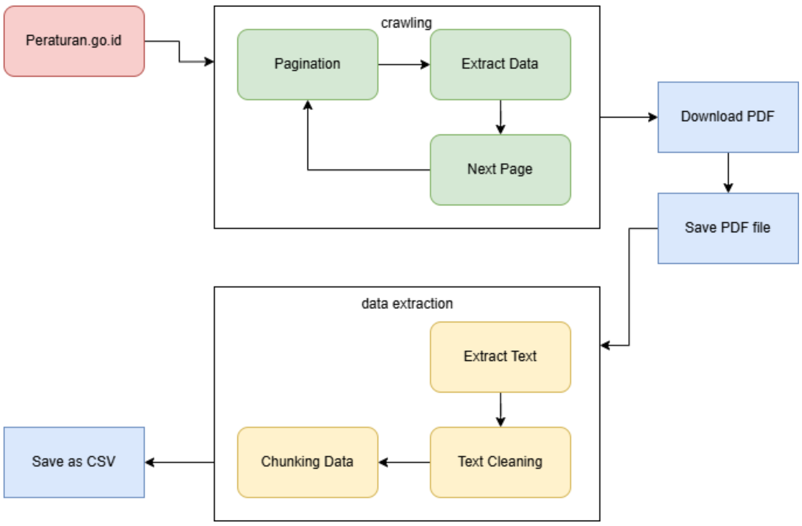

Figure 4. Dataset Collection Phase

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

The process flow for retrieving and extracting data from the URL https://peraturan.go.id is systematic. The process begins by accessing the site to perform crawling, which includes the pagination stage to navigate between pages, extract data to retrieve the displayed data, and next page to automatically continue to the next page until all data has been successfully collected.

After the data has been successfully retrieved, the process continues to downloading the documents in PDF format through the PDF download stage, then the file is saved locally (save PDF file). Next, the data extraction process is carried out on the downloaded PDF file.

This stage begins with extracting text to retrieve the text content from the PDF file, followed by text cleaning to remove irrelevant characters or elements from the text. After the text is cleaned, chunking data or dividing the data into smaller, more manageable parts is performed. The final step of the entire process is saving the extraction results into a CSV file through the save as CSV stage. The site has a total of 60.467 records, as shown in **Table 2**, making it a reliable and comprehensive source for the needs of this study.

Table 2. Number of Regulatory Data on Crawler System

Type of Regulation	Data Number
PERDA	19672
PERMEN	19143
PERBAN	6370
KEPPRES	5339
PP	4948
PERPRES	2585
UU	1889
INPRES	390
PERPPU	218
UUDRT	177
PENPRES	76
TAP MPR	41
UUD	1
Total	**60467**

Preprocessing

Preprocessing prepares raw data to suit the needs of transformer models such as DistilBERT so that the training and inference processes can run optimally. Preprocessing in this context includes text merging, label encoding, data splitting, and tokenization. Text merging is the process of combining several text columns in a dataset into one complete text column. Label encoding converts categorical data into a numerical format so that it can be processed. Prepare two labels for two different classification tasks (binary and multi-class), with status labels using binary and regulation type labels using multi-class. Data split divides the dataset into two (or more) parts, usually into training data and testing data. Split the dataset into two parts, namely training data (80%) and testing data (20%). The tokenization process uses DistilBERTTokenizer from the pre-trained distilbert-base-uncased model. Tokenization is an important process in NLP to convert raw text into a numerical format that can be processed by the model.

Model Training

The DistilBERT model was implemented and optimized using the hardware specifications listed in **Table 3**. This hardware combination supports model training efficiency, reduces the training time required, and provides optimal performance in classifying the status and type of legislation.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Table 3. Hardware Specifications

CPU	2 x Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Tesla P100 16GB
RAM	27GB
Storage	129GB available

In the DistilBERT model training process, the parameters listed in **Table 4** were used. The model used is a variant of DistilBERT, which is a distilled version of BERT, designed to achieve high performance with lighter computation. This combination of parameters supports fast training without compromising model accuracy and stability, ensuring that the DistilBERT model can be efficiently optimized for classifying the status and type of legal regulations.

Table 4. Parameters

Parameter	Value
Model variant	DistilBERT
Epoch	3
Per device train batch size	16
Per device eval batch size	16
Logging steps	10
Learning rate	5e-5
Weight decay	0.01

Model Evaluation

The evaluation results obtained after running the training on the model are presented in **Table 5**, which includes training loss and validation loss. This data provides an overview of the model's performance, which serves as the basis for further analysis in refining the model.

Table 5. Loss Evaluation Results

Epoch	1	2	3
Training Loss	0.130000	0.186200	0.146100
Validation Loss	0.154098	0.152164	0.148908

Training Loss measures how well the model learns from the training data. Validation Loss measures the model's performance on data it has never seen before (validation test). Ideally, both values should decrease consistently as the epoch increases, indicating that the model is learning well and not overfitting. In the first epoch, the training loss value is very small, around 0.13, while the validation loss is slightly higher at 0.15. This difference indicates that the model is learning and adapting to the training data very quickly, so its ability to recognize patterns in the training data is quite high. However, because the validation loss does not decrease significantly, this could also indicate that the model is starting to overfit, i.e., it is too well-suited to the training data but not yet optimal in dealing with new data. Entering the second epoch, the training loss value increased slightly to 0.18, while the validation loss decreased. The increase in training loss could be caused by regularization mechanisms, weight adjustments through techniques such as dropout, or the effect of a dynamic learning rate. This can be considered positive because the model may be trying to reduce overfitting and improve its ability to generalize the validation data. In the third epoch, the training loss decreased again and the validation loss also decreased slowly. This pattern indicates that the model is starting to achieve a balance between fit and generalization. The model is not only able to learn from the training data but also maintain stable performance on the validation data. This is a sign that the model is on the right track, neither overfitting nor underfitting the data used.

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

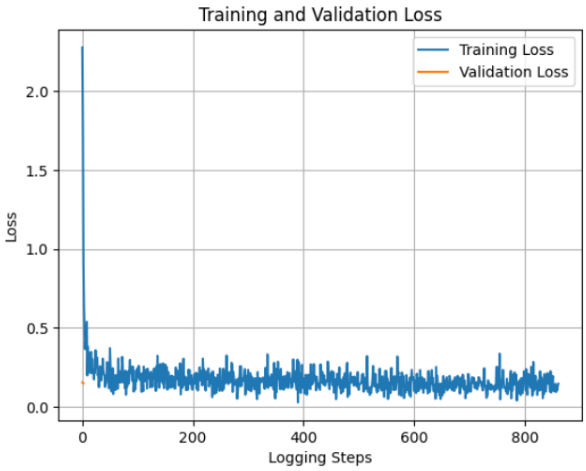

Figure 5. Training Loss and Validation Loss Graph

The Training Loss and Validation Loss graphs show how the loss values in the training data and validation data change during the model training process. The Training Loss and Validation Loss graphs are shown in **Figure 5**. The X-axis shows the logging steps, which is the number of iterations or batches in which the loss values are recorded, while the Y-axis shows the loss values. The two lines represent the training loss (blue) and validation loss (orange).

At the beginning of training, the training loss is very high, indicating that the model does not yet understand the data patterns well. However, in a short period, the training loss drops dramatically and then stabilizes at a low value, indicating that the model is learning quickly from the training data. Meanwhile, the validation loss is also low and stable, without any sharp spikes, which is a good sign that the model is not overfitting. The consistency and stability between the training loss and validation loss indicate that the model is learning well and has strong generalization capabilities for new data.

Classification evaluation is performed by calculating metrics such as Precision, Recall, F1-Score, and Support for each class. These metrics are provided in the form of a classification report table. The aim is to assess the model's performance in more detail than just accuracy. This evaluation provides insight into whether the model is suitable for use in the real world or needs further optimization. The results of the classification evaluation can be seen in **Figure 6**.

=== Status Classification ===

	precision	recall	f1-score	support
Tidak Berlaku	0.61	0.17	0.26	533
Berlaku	0.96	0.99	0.98	10966
accuracy			0.96	11499
macro avg	0.79	0.58	0.62	11499
weighted avg	0.94	0.96	0.94	11499

=== Jenis/Bentuk Peraturan Classification ===

	precision	recall	f1-score	support
INPRES	1.00	1.00	1.00	63
KEPPRES	1.00	1.00	1.00	1020
PENPRES	1.00	1.00	1.00	14
PERBAN	0.98	0.99	0.98	1152
PERDA	1.00	1.00	1.00	3724
PERMEN	1.00	0.99	1.00	3621
PERPPU	1.00	1.00	1.00	37
PERPRES	1.00	1.00	1.00	460
PP	1.00	1.00	1.00	1008
TAP MPR	1.00	1.00	1.00	6
UU	1.00	1.00	1.00	364
UUDRT	1.00	1.00	1.00	30
accuracy			1.00	11499
macro avg	1.00	1.00	1.00	11499
weighted avg	1.00	1.00	1.00	11499

Figure 6. Classification Evaluation Results

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

This evaluation uses common metrics in classification, namely precision, recall, f1-score, and accuracy. The results show the model's performance in predicting labels for each category available in the test data, which consists of 11,499 data points.

In the Status Classification section, there are two classes, namely “Not Applicable” and “Applicable.” The ‘Applicable’ class performed very well with a precision of 0.96, a recall of 0.99, and an f1-score of 0.98. This means that the model is very accurate in recognizing and predicting data with the status “Applicable.” The overall accuracy remained high at 0.96.

In the Type/Form of Regulation section, the model's performance is impressive. All classes (such as INPRES, KEPRES, PERMEN, UU, etc.) obtained near-perfect precision, recall, and f1-score values of 1.00. Even for labels with a small amount of data, such as TAP MPR and PENPRES, the model was still able to predict accurately. This indicates that the model can consistently and accurately distinguish between types of regulations. The accuracy, macro average, and weighted average values all also achieve a perfect score of 1.00, indicating no classification errors for this label in the test data.

A confusion matrix is a table that shows the number of correct and incorrect predictions made by a model for each class. Each cell in the matrix represents the number of predictions for a particular combination of actual labels and predicted labels. The Confusion Matrix can be seen in **Figure 7**.

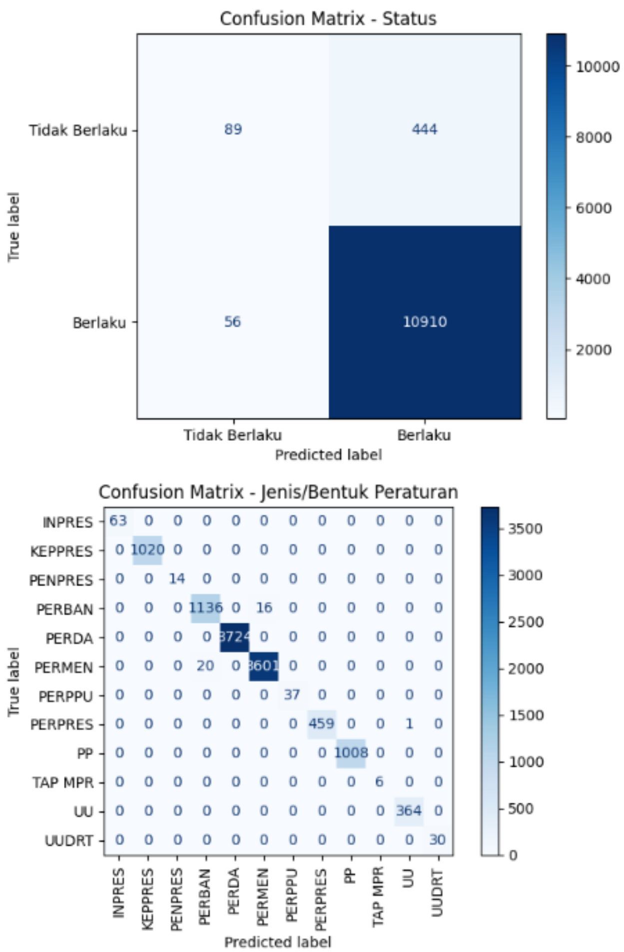

Figure 7. Confusion Matrix Status and Type of Regulation

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In the upper confusion matrix section titled “Confusion Matrix - Status,” it can be seen that the model classifies two classes, namely Applicable and Not Applicable. The results show that the model is very good at classifying documents with Applicable status, with 10,910 correct predictions and only 56 misclassified as Not Applicable.
In the lower section, “Confusion Matrix - Type/Form of Regulation,” the model's classification performance for various types or forms of regulations such as INPRES, KEPPRES, PERDA, UU, and so on is displayed. Here, the model's performance is generally quite good, as evidenced by the dominance of numbers on the main diagonal of the matrix (which indicates correct predictions). Classification errors between types of regulations appear to be minimal, as seen from the large number of zeros outside the diagonal. This indicates that the model has high accuracy and good precision in distinguishing types of legal regulations.
The use of classification predictions refers to the application of model results in a real-world context. Here, the evaluated model output is used to make decisions. The results of single predictions of regulation names based on status and type of regulation are shown in **Figure 8**.

```
Nama Peraturan: Peraturan Presiden Nomor 54 Tahun 2020 Tentang Penanganan COVID-19
Prediksi Status: Berlaku (Confidence: 1.0)
Prediksi Jenis: PERPRES (Confidence: 1.0)

Nama Peraturan: Undang-Undang Nomor 23 Tahun 2014 Tentang Pemerintahan Daerah
Prediksi Status: Berlaku (Confidence: 0.99)
Prediksi Jenis: UU (Confidence: 1.0)

Nama Peraturan: Peraturan Menteri Kesehatan Nomor 9 Tahun 2021 Tentang Karantina Wilayah
Prediksi Status: Berlaku (Confidence: 1.0)
Prediksi Jenis: PERMEN (Confidence: 1.0)

Nama Peraturan: Peraturan Daerah Nomor 3 Tahun 2019 Tentang Ketertiban Umum
Prediksi Status: Berlaku (Confidence: 1.0)
Prediksi Jenis: PERDA (Confidence: 1.0)
```

Figure 8. Single Prediction Results for Regulation Names Based on Status and Type of Regulation

These results indicate that the model has high accuracy and confidence in classifying the type and status of regulations based solely on the title text. This indicates that the model has learned a good representation of the structure and patterns of language in official regulatory documents in Indonesia. The results of the prediction of the sound column of verses based on the status and type of regulation are shown in **Figure 9**.

```
                                        Title  \
0  Peraturan Pemerintah Pengganti Undang-Undang N...
1  Peraturan Pemerintah Pengganti Undang-Undang N...
2  Peraturan Pemerintah Pengganti Undang-Undang N...
3  Peraturan Pemerintah Pengganti Undang-Undang N...
4  Peraturan Pemerintah Pengganti Undang-Undang N...

                                   BUNYI_AYAT Prediksi Status  \
0  Pengurus semua perusahaan/badan yang berusaha/...         Berlaku
1  Menteri Urusan Pendapatan, Pembiayaan dan Peng...         Berlaku
2  Surat-paksa dilaksanakan sesuai dengan ketentu...         Berlaku
3  Barangsiapa menguasai dana investasi yang pemi...         Berlaku
4  Surat-paksa dilaksanakan sesuai dengan ketentu...         Berlaku

   Confidence Status Prediksi Jenis  Confidence Jenis
0           0.999310         PERBAN          0.935373
1           0.873123         PERMEN          0.999670
2           0.986547         PERBAN          0.997960
3           0.996541         PERBAN          0.996129
4           0.986547         PERBAN          0.997960
```

Figure 9. Predicted Results of Verse Sound Columns Based on Status and Type of Regulation

Journal of Information Systems Engineering and Management
2025, 10(52s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This model predicts the status of regulations and the type or form of regulations, providing a confidence score for each prediction. The model performs very well in predicting the status and type of regulations based on a combination of titles and article content. A high confidence score indicates that the model is quite confident in its classification. The results of single sound predictions based on the status and type of regulations are shown in **Figure 10**.

```
Bunyi Ayat: Wakil Menteri sebagaimana dimaksud dalam Pasal 2 ayat (1) juga merupakan Wakil Kepala Badan Koordinasi Penanaman Modal.
Prediksi Status: Berlaku
Prediksi Jenis/Bentuk Peraturan: PERMEN
```

Figure 10. Single Prediction Results of Verse Sounds Based on Status and Type of Regulation

Based on the content of this paragraph, the model predicts that the status of the regulation is “Applicable” The model appears to have successfully captured the bureaucratic context of the paragraph to classify the form of regulation. These results indicate that the model has the ability to understand the context of regulatory texts to classify the form and status of regulations well.

CONCLUSION

The DistilBERT model successfully classified two labels effectively. The model showed excellent performance. For Regulation Status classification, the model achieved an accuracy of 96%, precision of 94%, recall of 96%, and f1-score of 94%, reflecting the model's accuracy and consistency in identifying the “Applicable” and “Not Applicable” statuses overall. Meanwhile, in the classification of Regulation Type/Form, the model showed near-perfect results with accuracy, precision, recall, and f1-score of 100%, indicating that the model can classify various types of regulations with high accuracy. In general, the high weighted average value indicates that the model has worked optimally on the existing data distribution.

Based on the research results, it is recommended to balance the data distribution in the Regulation Status label, especially the Not Applicable class, to reduce bias towards the majority class and improve model performance on the recall and f1-score metrics. Techniques such as oversampling, undersampling, or SMOTE can be considered. In addition, exploration of alternative transformer models such as IndoBERT, RoBERTa, or BigBird is necessary to compare accuracy and efficiency, given each model's ability to handle long legal texts in Indonesian. Further testing may also include original legal documents in PDF format or OCR results to assess the effectiveness of the model in real-world scenarios.

Refrences

[1] Akca, O., Bayrak, G., Issifu, A. M., & Ganiz, M. C. (2022). Traditional Machine Learning and Deep Learning-based Text Classification for Turkish Law Documents using Transformers and Domain Adaptation. *16th International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2022*. https://doi.org/10.1109/INISTA55318.2022.9894051

[2] Alexopoulos, C., Loukis, E., & Virkar, S. (2024). Architecture and Evaluation of an Advanced Legal Information Platform—Enhancing Productivity of Modern Legal Work. *Journal of the Knowledge Economy*, *15*(2). https://doi.org/10.1007/s13132-023-01415-5

[3] Azhar, N., & Latif, S. (2022). Roman Urdu Sentiment Analysis Using Pre-trained DistilBERT and XLNet. *Proceedings - 2022 5th International Conference of Women in Data Science at Prince Sultan University, WiDS-PSU 2022*. https://doi.org/10.1109/WiDS-PSU54548.2022.00027

[4] Bambroo, P., & Awasthi, A. (2021). LegalDB: Long distilbert for legal document classification. *Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing,*

Communications and Sustainable Technologies, ICAECT 2021. https://doi.org/10.1109/ICAECT49130.2021.9392558

[5] Berfu B, Ali H, Arzucan¨, A., & Arzucan¨ozgür, A. A. (2020). Analyzing ELMo and DistilBERT on Socio-political News Classification. *Language Resources and Evaluation Conference*.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

[7] Dong, J. (2024). Natural Language Processing Pretraining Language Model for Computer Intelligent Recognition Technology. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(8). https://doi.org/10.1145/3605210

[8] Fang, Z., He, Y., & Procter, R. (2023). *CWTM: Leveraging Contextualized Word Embeddings from BERT for Neural Topic Modeling*. http://arxiv.org/abs/2305.09329

[9] Firma, Z., Pusat, A., Dan, P., Perkara, P., Konstitusi, M., Indonesia, R., Medan, J., Nomor, M. B., Pusat, J., Yulistyaputri, R., & Penelitian, P. (2019). *ROMANTISME SISTEM HUKUM DI INDONESIA : KAJIAN ATAS KONSTRIBUSI HUKUM ADAT DAN HUKUM ISLAM TERHADAP PEMBANGUNAN HUKUM DI INDONESIA (The Romanticism of Legal Systems in Indonesia: The Study of The Constribution of Islamic Law And Islamic Law for Legal Development In Indonesia)*.

[10] Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic Similarity Metrics for Evaluating Source Code Summarization. *IEEE International Conference on Program Comprehension*, *2022-March*, 36–47. https://doi.org/10.1145/

I. Turc, M.-W. C. K. L. and K. T. (2019). *WELL-READ STUDENTS LEARN BETTER: ON THE IMPORTANCE OF PRE-TRAINING COMPACT MODELS*.

[11] Jojoa, M., Eftekhar, P., Nowrouzi-Kia, B., & Garcia-Zapirain, B. (2024). Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. *AI and Society*, *39*(3). https://doi.org/10.1007/s00146-022-01594-w

[12] Joshy, A., & Sundar, S. (2022). Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa. *2022 IEEE International Power and Renewable Energy Conference, IPRECON 2022*. https://doi.org/10.1109/IPRECON55716.2022.10059542

[13] Kumar, A., Saumya, S., & Singh, J. P. (2020). NITP-AI-NLP@HASOC-FIRE2020: Fine tuned BERT for the Hate Speech and Offensive Content identification from social media. *CEUR Workshop Proceedings*, *2826*.

[14] Kumari, V., Keshari, S., Sharma, Y., & Goel, L. (2022). Context-Based Question Answering System with Suggested Questions. *Proceedings of the Confluence 2022 - 12th International Conference on Cloud Computing, Data Science and Engineering*. https://doi.org/10.1109/Confluence52989.2022.9734207

[15] Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., Goin, M., & Alistarh, D. (2022). The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. https://doi.org/10.18653/v1/2022.emnlp-main.279

[16] Lê, J. K., & Schmid, T. (2022). The Practice of Innovating Research Methods. *Organizational Research Methods*, *25*(2). https://doi.org/10.1177/1094428120935498

[17] Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00782-9

[18] Liu, Z., Xu, Z., Jin, J., Shen, Z., & Darrell, T. (2023). Dropout Reduces Underfitting. *Proceedings of Machine Learning Research*, *202*.

[19] Maheshwari, V., Dutta, K., & Kushwaha, A. (2023). Analysis of Different Adversarial Attacks on Various NLP SoTA Models. *Proceedings - 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications, CIISCA 2023*. https://doi.org/10.1109/CIISCA59740.2023.00049

[20] Mollá, D. (2022). Query-focused Extractive Summarisation for Biomedical and COVID-19 Complex Question Answering. *CEUR Workshop Proceedings*, *3180*.

[21] Petrakova, M. S. (2022). Systematization of legislation as a way to ensure the effectiveness of legal regulation: theoretical and legal aspects. *Courier of Kutafin Moscow State Law University (MSAL))*, *1*(11). https://doi.org/10.17803/2311-5998.2021.87.11.233-240

[22] Quevedo, E., Rahman, M., Cerny, T., Rivas, P., & Bejarano, G. (2022). *Study of Question Answering on Legal Software Document using BERT based models*. https://doi.org/10.52591/lxai202207103

[23] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108

[24] Simanjuntak, E. (2019). Peran Yurisprudensi dalam Sistem Hukum di Indonesia. *Jurnal Konstitusi*, *16*(1). https://doi.org/10.31078/jk1615

[25] Xie, X., Xie, M., Moshayedi, A. J., & Noori Skandari, M. H. (2022). A Hybrid Improved Neural Networks Algorithm Based on L2 and Dropout Regularization. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/8220453