Jurnal Informatika: Jurnal pengembangan IT
Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.8795

Comparison of IndoBERT and Bi-LSTM Models for Indonesian Law Violation Text Classification

Made Wahyu Adwitya Pramana [1], Desy Purnami Singgih Putri [2], I Ketut Adi Purnawan [3]
Teknologi Informasi, Universitas Udayana, Jl. Raya Kampus UNUD Bukit Jimbaran, Bali, 80361, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-05-21
Revised 2025-08-29
Accepted 2025-08-30

Corresponding Author:
Made Wahyu Adwitya Pramana
Email:
mdwahyu.pramana@gmail.com

Abstract *– Legal violations in Indonesia, particularly those under the Criminal Code (KUHP) and the Information and Electronic Transactions Law (UU ITE), are often difficult for the general public to interpret due to the complexity of legal language and article structures. This research aims to build a multilabel classification model that can automatically identify relevant legal articles from user-provided case descriptions. Two models were developed and compared: Bidirectional Long Short-Term Memory (Bi-LSTM) and IndoBERT. Using a manually labeled dataset, both models were evaluated through accuracy, F1-score, and Hamming Loss metrics, as well as 5-fold cross-validation. The results showed that IndoBERT outperformed Bi-LSTM with an average accuracy of 97% and a Hamming Loss of 0.027. However, t-test analysis revealed no statistically significant difference in F1-scores, indicating that both models have comparable effectiveness in capturing multiple labels. A confusion matrix analysis further identified patterns of misclassification in semantically similar articles. This study demonstrates the potential of NLP and deep learning to support legal awareness and provide the public with easier access to legal information.*

Keywords: *Bi-LSTM; IndoBER; KUHP; Text Mining; UU ITE.*

Abstrak *– Pelanggaran hukum di Indonesia, khususnya yang tercakup dalam Kitab Undang-Undang Hukum Pidana (KUHP) dan Undang-Undang Informasi dan Transaksi Elektronik (UU ITE), sering kali sulit dipahami oleh masyarakat umum karena kompleksitas bahasa hukum dan struktur pasalnya. Penelitian ini bertujuan untuk membangun model klasifikasi multilabel yang dapat secara otomatis mengidentifikasi pasal hukum yang relevan berdasarkan deskripsi kasus yang diberikan oleh pengguna. Dua model dikembangkan dan dibandingkan, yaitu Bidirectional Long Short-Term Memory (Bi-LSTM) dan IndoBERT. Menggunakan dataset berlabel secara manual, kedua model dievaluasi menggunakan metrik akurasi, F1-score, dan Hamming Loss, serta validasi silang 5-fold. Hasil menunjukkan bahwa model IndoBERT memiliki performa lebih baik dibandingkan Bi-LSTM dengan rata-rata akurasi sebesar 97% dan Hamming Loss sebesar 0,027. Namun, analisis uji t menunjukkan bahwa tidak terdapat perbedaan yang signifikan secara statistik pada nilai F1-score, yang mengindikasikan bahwa kedua model memiliki efektivitas yang sebanding dalam menangkap beberapa label sekaligus. Analisis confusion matrix juga mengungkapkan adanya pola salah klasifikasi pada pasal-pasal yang memiliki kemiripan makna. Penelitian ini menunjukkan potensi NLP dan deep learning dalam mendukung kesadaran hukum dan memberikan akses informasi hukum yang lebih mudah bagi masyarakat.*

Kata Kunci*: UU ITE, KUHP, Text Mining, Bi-LSTM, IndoBERT*

I. Introduction

The rapid development of information and communication technology has introduced new challenges to law enforcement in Indonesia, particularly concerning violations occurring in the digital realm. The Law on Electronic Information and Transactions (UU ITE) serves as the main legal framework for regulating various criminal acts arising from the misuse of technology, such as the spread of false information (hoaxes), hate speech, cyberbullying, and online gambling [1][2]. On the other hand, the Indonesian Criminal Code (KUHP) continues to play a vital role in addressing conventional crimes, including defamation and insult, which often overlap with provisions in the UU ITE.

The complexity and overlap between the articles in UU ITE and the Penal Code often lead to varied interpretations in legal practice, which makes identifying and classifying violations a challenging task for law enforcement officers [3]. Additionally, manually searching for relevant legal statutes in case files or digital records is not only time-consuming but also susceptible to mistakes. This situation highlights the increasing need for automated tools powered by artificial intelligence to facilitate faster and more accurate legal classifications.

Natural Language Processing (NLP) combined with deep learning has shown great potential for automating text classification tasks. NLP workflows typically involve cleaning the text, splitting it into tokens, removing common stopwords, and applying stemming, before converting the text into numerical vector forms suitable for model input [4]. Models such as Bidirectional Long Short-Term Memory (Bi-LSTM) use recurrent neural networks that analyze sequences forwards and backwards, which helps capture richer contextual details

[5]. IndoBERT, a transformer model pretrained on extensive Indonesian language data, uses attention mechanisms to deeply grasp the context of words and sentences. Both of these models are well-suited for multilabel classification, as they can effectively recognize semantic and contextual subtleties within Indonesian texts.

Prior studies have investigated the application of deep learning for categorizing legal violations, with a particular focus on Indonesia's UU ITE. For instance, a 2023 study utilized LSTM and Bi-LSTM architectures to classify criminal offenses based on Twitter data [6]. This research collected 17,384 tweets through the Twitter API and earlier works. The models were evaluated in two conditions: with and without dropout layers intended to reduce overfitting. The results showed that the Bi-LSTM model with dropout performed best, achieving an F1-score of 0.9301 and an accuracy of 0.9807, outperforming the standard LSTM. The primary categories analyzed included pornography, hoaxes, cyberbullying, and hate speech, which demonstrated promising applications for supporting law enforcement under UU ITE.

Another relevant study investigated toxic comment detection in social media using pre-trained transformer models, including IndoBERT, mBERT, and IndoRoBERTa Small [7]. IndoBERT, trained on large-scale Indonesian corpora, achieved the highest F1-score of 0.8897, outperforming the other models. This work demonstrated the effectiveness of transformer-based models in multi-label classification tasks involving the Indonesian language. While the objective of this study centered on identifying toxic comments—such as hate speech, radicalism, and pornography—the methodology closely aligns with the current research, particularly in the utilization of IndoBERT for Indonesian-language text classification.

In another study, researchers used Naive Bayes to classify legal violations under UU ITE based on textual descriptions from court decisions [8]. The study focused on articles 27 and 28 of the UU ITE and employed a dataset of 245 violation descriptions sourced from Supreme Court rulings. The Naive Bayes classifier achieved an accuracy of 80% using 196 training samples and 49 test samples. While the approach demonstrated satisfactory results, it was limited in scope, both in the variety of articles considered and in the classical machine learning technique employed. In contrast, the present study aims to address a broader classification task covering both UU ITE and KUHP violations using more advanced deep learning techniques.

Although earlier studies have shown promising results, there are still some limitations. Most research tends to concentrate either on informal user-generated content or on a narrow set of UU ITE articles, with little attention paid to integrating provisions from the KUHP. Additionally, few efforts have applied deep learning techniques to multilabel classification of legal violations using formal, descriptive case texts, which better represent actual legal documents.

To fill these gaps, this study takes a comprehensive approach by employing both Bidirectional Long Short-Term Memory (Bi-LSTM) and IndoBERT models for multilabel classification of legal articles. Bi-LSTM is selected due to its ability to capture dependencies across long text sequences, making it well-suited for modeling detailed legal case narratives [9]. Meanwhile, IndoBERT benefits from extensive pretraining on Indonesian language data, which helps it understand complex legal terminology and context [10]. These methods complement each other, combining Bi-LSTM's sequence analysis with IndoBERT's deep contextual understanding, offering a more effective solution for handling Indonesia's intricate legal language. The dataset used in this study originates from web scraping legal case descriptions on Hukumonline and SIPP (Sistem Informasi Penelusuran Perkara), ensuring that the data reflect authentic and formal legal sources.

This research contributes a novel blend of IndoBERT and Bi-LSTM for multilabel classification across both UU ITE and KUHP provisions. Unlike previous work focusing on social media data or limited article subsets, this study uses formal case descriptions and implements a rigorous comparison supported by statistical tests and web-based deployment. Beyond assessing model performance, it also explores patterns of misclassification to identify common errors. The findings aim to support the development of more reliable and interpretable legal NLP tools, improving public access to legal information and contributing to greater legal awareness and justice through technology.

Commented [LA1]: Dijadikan 1 paragraf sebagai kontribusi

The classification approach allows predicting multiple legal articles simultaneously by using sigmoid activation with class-specific thresholds. Performance is assessed with macro and micro F1-scores to balance class-level and overall accuracy, along with Hamming Loss to measure multilabel misclassifications. The models evaluated include Bi-LSTM, which captures sequential patterns, and IndoBERT, which provides context-aware word representations specifically for Indonesian legal language.

II. Methodology

The research was carried out in multiple steps. It began with gathering the data, followed by cleaning data and preparing it for analysis. Next, the IndoBERT and Bi-LSTM models were trained using this processed data. The last step focused on assessing the models' performance by applying evaluation measures suitable for multilabel classification, including accuracy, Hamming loss, and F1 score.

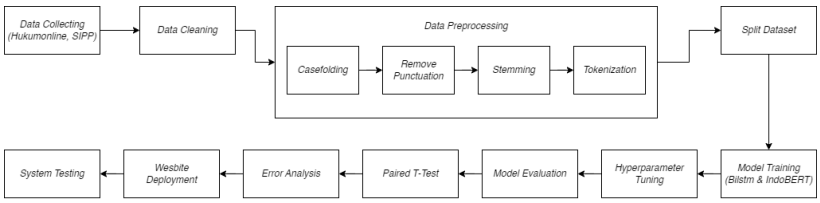

Figure 1. Steps of Research

Figure 1 shows the steps that used to build this research. The study started by collecting data from two main sources, there are Hukumonline and SIPP. After gathering the data, cleaning data was carried out to remove duplicates, irrelevant items, and spelling mistakes. The next step involved text preprocessing, which included converting all text to lowercase, breaking it into tokens, removing common stop words, and applying stemming to prepare the data for modelling step. The dataset was then split into training and testing. The next step is the Bi-LSTM and IndoBERT models were customized and trained, with hyperparameters adjusted to improve their performance. Model results were evaluated using metrics like accuracy, macro and micro F1 scores, Hamming Loss, and Precision-Recall curves. To check if there are differences between model performances, a paired statistical test was performed. Error analysis was also done for each class to uncover common mistakes in predictions. Finally, the model with the best results was implemented into a web application built with Streamlit, and system testing that including usability and black box tests to confirm it met user needs.

A. Data Preparation

This study utilizes a dataset of legal cases related to the Electronic Information and Transactions Law (UU ITE) and the Criminal Code (KUHP), collected through web scraping from Hukumonline and SIPP. To ensure adequate data representation for training and avoid class imbalance, only articles (articles) with the highest frequency of occurrence in the dataset were included in this study. This selection strategy was implemented as a deliberate scope limitation to focus the model on the most frequently encountered violations, thus enabling the development of a more robust and reliable classifier within the available data constraints. Expanding the model to include legal articles with lower frequency of occurrence remains a potential direction for future research.

TABEL 1
DATESET DESCRIPTION

Feature	Description
text	This feature represents textual descriptions of legal violation cases that including the Electronic Information and Transactions Law (UU ITE) and the Criminal Code (KUHP).
pasal	This feature represents the relevant articles of the ITE Law or the Criminal Code that are relevant to each case.

TABEL 2
DEATESET DISTRIBUTION

Feature	Description	Percentage
Hukumonline	7056	51%
SIPP	6642	49%

The dataset used in this study was collected from two primary sources, there are Hukumonline and SIPP.There are 7,056 data samples (51%) were obtained from Hukumonline, and 6,642 data samples (49%) were collected from SIPP. This distribution reflects a relatively balanced composition of legal case data from both sources, ensuring a diverse representation of legal issues related to the ITE Law and the Criminal Code.

B. Data Cleaning

The data cleaning phase is the phase to ensuring the consistency and reliability of the dataset [11]. This phase involves several tasks, there are extracting relevant text, removing null value, reducing noise, filtering the dataset, and combining data from multiple sources. These steps help prepare the data for further preprocessing and modeling by removing irrelevant or inconsistent entries and combining multiple datasets into a unified format.

TABEL 3
PASAL DISTRIBUTION

Feature	Counts
Pasal 27 Ayat (1) UU ITE	1172
Pasal 27 Ayat (2) UU ITE	1039
Pasal 27 Ayat (3) UU ITE	790
Pasal 303 KUHP	504
Pasal 363 KUHP	407
Pasal 368 KUHP	291

The table above shows how the legal articles are distributed in the dataset. Pasal 27 Ayat (1) UU ITE appears most frequently with 1,172 examples, followed by Pasal 27 Ayat (2) with 1,039 instances and Pasal 27 Ayat (3) with 790. Articles from the KUHP appear less often, such as Pasal 368 with only 291 cases. This unequal distribution can cause challenges during training because the model might struggle to classify less common categories. Instead of balancing the dataset by oversampling or undersampling, this research applies class weighting during training. This method adjusts the importance of each class in the loss function, giving higher weight to minority classes to prevent the model from being biased toward majority classes.

C. Data Preprocessing

Preprocessing transforms raw and inconsistent text data into a structured form suitable for machine learning algorithms [12]. In this study, the dataset underwent several standard preprocessing steps to prepare it for training and improve data quality [13]. Several preprocessing steps were applied in this phase, including case folding, punctuation removal, stopwords removal, stemming, and tokenization to split the text into individual tokens for sequence modeling.

When preparing data for sequence models, the maximum sequence length was set differently depending on the model. For Bi-LSTM, the limit was based on the average sequence length from the dataset, with sequences padded or truncated as needed. IndoBERT used a fixed maximum length of 128 tokens, following its default setup. These preprocessing actions help simplify the text, reduce noise, and maintain consistent input formats for the models.

D. Bi-LSTM

This research uses the Bidirectional Long Short-Term Memory (Bi-LSTM) model, a type of deep learning method for classification. Bi-LSTM enhances the usual LSTM by processing sequences both forward and backward, allowing it to capture context from before and after each word [14]. This bidirectional processing improves understanding of the semantic structure within legal case descriptions. The input is first converted into dense vector embeddings representing the words, which the Bi-LSTM layers then analyze to find patterns relevant to classifying legal violations under UU ITE and KUHP. The final output layer employs a sigmoid activation function for multilabel classification, enabling the model to assign multiple legal articles to a single input. The model's performance is measured using metrics such as accuracy, Hamming loss, and F1-score.

E. IndoBERT

IndoBERT, based on the BERT architecture, is specifically trained for the Indonesian language. BERT models use transformers to understand a word's context by reading text in both directions, left to right and right to left [2]. This bidirectional approach allows for a deep grasp of sentence meaning. In this study, the IndoBERT model from the indobenchmark/IndoBERT-base-p2 checkpoint was used to improve understanding of Indonesian legal texts. The model was pretrained with two main tasks: Masked Language Modeling (MLM), which trains the model to predict missing words in sentences, and Next Sentence Prediction (NSP), which teaches it to recognize relationships between sentences

Commented [LA2]: Gunakan kalimat aktif untuk jurnal berbahasa inggris

F. Model Evaluation

The evaluation phase is aimed to compare the performance between the models that have been trained. The model that compared in this research is model based on Bi-LSTM and the other on IndoBERT. Both models

were tested on the same dataset to ensure a fair comparison. The following methods were used to evaluate their effectiveness.

1) *Classification Report*: This report summarizes key metrics derived from the confusion matrix, which tracks true positives, true negatives, false positives, and false negatives [15]. These metrics help to show the strengths and weaknesses of every model that have been trained, which is especially important for handling the complexity of multilabel legal classification tasks.

2) Cross Validation: To improve reliability and assess how well the models generalize, the study applied K-Fold Cross Validation. The dataset was split into k equal parts or folds, and the model was trained and validated k times—each time using a different fold as the validation set and the others for training. This method reduces overfitting and provides a stronger estimate of the model's real-world accuracy [16]. Metrics like accuracy, Hamming loss, and F1-score were recorded for each fold and analyzed to compare the two models. Using cross-validation strengthens confidence in the results and offers a balanced basis for comparison.

3) *Confusion Matrix*: The confusion matrix was used to break down the correct and incorrect predictions by comparing predicted labels with actual labels for each class [15]. I For multilabel classification, a separate matrix was created for each label, allowing detailed analysis of mistakes such as false positives and negatives. From these, precision, recall, and F1-score were calculated to measure how well the model recognized each class. The confusion matrix also helped identify which legal categories were more challenging to classify accurately, guiding further improvement.

G. Website Implementation

To make the classification model easily accessible, a web application was developed using Streamlit, a Python framework for building interactive machine learning tools. This app allows users to enter legal case descriptions in natural language. The input is processed and passed through either the Bi-LSTM or IndoBERT model to predict related legal articles from UU ITE or the Indonesian Penal Code (KUHP). The system includes preprocessing steps, model inference, and a clean interface that displays the results clearly. This setup aims to provide a user-friendly tool so the public can obtain legal insights quickly and easily.

H. System Testing

System testing ensured the application functioned as intended. Testing focused on outputs using black-box methods, where internal code was not examined. Each major feature was tested with inputs representing typical user actions, and outputs were checked to match expected results. Tests ran in a local environment accessible only to the tester, without external users or live deployment. This phase validated the core functions before the system's wider release.

III. Result and Discussion

This section discusses the evaluation of the models' performance, focusing on both IndoBERT and Bi-LSTM. The results are compared to determine which model delivers better outcomes based on the dataset used.

A. Model Performance

This phase shows the performance comparison between the Bi-LSTM and IndoBERT models with several experimental scenarios. Each model was tested with multiple configurations to see how changes in design and training affected their ability to classify legal texts. The goal was to find the best setup for each model. After selecting the top configuration for both, K-Fold Cross Validation was applied to evaluate their stability and reliability across different portions of the data. The findings provide insights into the strengths and weaknesses of each model when applied to Indonesian legal text classification.

TABEL 4
BI-LSTM MODEL ARCHITECTURE

Layer	Model 1	Model 2	Model 3
Embedding	300 dim, trainable		
Spatial Dropout	–	0.2	0.2
Batch Normalization	2x	3x	3x
BiLSTM 1	64, return seq, L2=0.01	128, return seq, L2=0.001	128, return seq

Dropout 1	0.3	0.3	0.3
BiLSTM 2	32	64, L2 & recurrent L2	64
Dropout 2	0.3	0.3	0.3
Dense 1	32, ReLU, L2=0.01	128, ReLU, L2=0.001	128, ReLU
Dense 2	–	64, ReLU, L2=0.001	64, ReLU
Dropout (Dense)	0.3	0.3 (2x)	0.3 (2x)
Output Layer	sigmoid, multilabel		

TABEL 5
BI-LSTM MODEL PERFORMANCE

Model	Sc.	Metrics		
		Accuracy (%)	**Hamming Loss**	**F1-Score (%)**
Bi-LSTM	I	88,66%	0.0322	89,97%
	II	88,66%	0.0325	89,62%
	III	**90,89%**	**0.0280**	**90,95%**

Table 4 shows the Bi-LSTM models performance, three experimental scenarios were tested to determine the optimal setting for multilabel classification of legal articles. Each model trained for 50 epochs to ensure sufficient learning time. In the first two scenarios, the model reached the same accuracy of 88.66%, with minor differences in Hamming Loss (0.0322 and 0.0325) and F1-Scores (89.97% and 89.62%). Scenario three showed a clear improvement, with the model achieving 90.89% accuracy, lowering Hamming Loss to 0.0280, and increasing the F1-Score to 90.95%. This indicates that the third scenario best captures relevant patterns in legal texts, improving the model's ability to correctly assign multiple legal articles without overfitting. Therefore, this configuration was chosen for further testing, including cross-validation and comparison with IndoBERT.

TABEL 6
INDOBERT MODEL ARCHITECTURE

Component	**Model Indobert-uncased**	**Model Indobert-base-p2**
Model Type	IndoBERT Uncased	IndoBERT Base + Dropout
Layer Freeze	No	No
Dropout	No	Yes (Dropout 0.3)
Learning Rate	0.00002	0.00002
Optimizer	AdamW	AdamW
Output Layer	BertForSequenceClassification	Linear layer custom

TABEL 5
INDOBERT MODEL PERFORMANCE

Feature	Sc.	Metrics		
		Accuracy (%)	**Hamming Loss**	**F1-Score (%)**
Indobert-base-p2	I	89,81%	0.0298	91,06%
	II	**90,29%**	**0.0290**	**91,30%**
Indobert-uncased	I	86,81%	0.0404	87,75%

Table 5 show IndoBERT models performance, two pretrained versions IndoBERT-base-p2 and IndoBERT-uncased—were evaluated with different settings. Thanks to pretrained language models, fine-tuning was limited to only 5 epochs, sufficient for adaptation to the classification task. IndoBERT-base-p2 in Scenario II performed best, with the highest accuracy of 90.29%, a Hamming Loss of 0.0290, and an F1-Score of 91.30%. Scenario I with the same model also showed strong results (89.81% accuracy and 91.06% F1-Score), reflecting

consistent performance. On the other hand, IndoBERT-uncased in Scenario I gave the weakest outcomes among the IndoBERT options, posting 86.81% accuracy, a higher Hamming Loss of 0.0404, and an F1-Score of 87.75%. These results suggest that IndoBERT-base-p2, particularly in Scenario II, is more effective in understanding the context of Indonesian legal texts, making it the preferred choice for further cross-validation and head-to-head comparison with Bi-LSTM.

B. Cross Validation

To evaluate how well the models generalize and maintain consistency, we performed 5-fold cross-validation. This technique divides the dataset into five parts, trains the model on four parts, and validates it on the remaining part. This process repeats five times so that each subset is used once for validation. By averaging the results across all folds, we obtain a more reliable measure of model performance, reducing potential bias from a single data split. Both the Bi-LSTM and IndoBERT models were evaluated this way, with IndoBERT consistently achieving higher accuracy and lower Hamming Loss than Bi-LSTM on every fold.

TABEL 6
CROSS VALIDATION

Model	**Fold Accuracy**					**Avg ± SD**	**Avg ± SD**
	1	2	3	4	5	**Acc.**	**Hamming Loss**
Bi-LSTM	89%	91%	93%	90%	91%	90% ± 1.48%	90% ± 1.48%
IndoBERT	96%	97%	97%	97%	97%	97% ± 0.73%	97% ± 0.73%

Looking closer at the cross-validation results from table 6, the Bi-LSTM model reached an average accuracy of 90% and an average Hamming Loss of 0.029 across all folds. This evaluation helps highlight the models' strengths and weaknesses more clearly than accuracy alone, offering useful guidance for future improvements in classification. Meanwhile, the IndoBERT model consistently outperformed Bi-LSTM, achieving an average accuracy of 97% and a lower average Hamming Loss of 0.027, indicating more precise multilabel classification performance.

To assess whether these differences were statistically significant, a paired t-test was conducted across the same folds for accuracy, Hamming Loss, and F1 score. Results revealed significant differences in both training accuracy ($t(4) = 27.91$, $p < 0.00001$) and validation accuracy ($t(4) = 13.45$, $p < 0.00001$). The effect size for accuracy, measured using Cohen's d, was 2.94 (large effect). However, differences in F1 score ($t(4) = 1.39$, $p = 0.195$, $d = 0.32$) and Hamming Loss ($t(4) = -1.32$, $p = 0.217$, $d = 0.29$) were not statistically significant, suggesting comparable performance in terms of precision–recall balance and multilabel prediction errors.

In summary, IndoBERT shows a statistically significant and large improvement in overall accuracy compared to Bi-LSTM, while their F1 score and error rates (Hamming Loss) remain similar across folds. This suggests IndoBERT's advantage lies primarily in correctly classifying more instances overall, while both models handle multilabel classification errors similarly.

C. Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's predictions by comparing true labels against predicted labels for each class. It helps identify specific types of errors such as false positives and false negatives, offering insights into which legal articles are more frequently misclassified. This analysis supports a deeper understanding of model strengths and weaknesses beyond overall accuracy, guiding further improvements in classification performance.

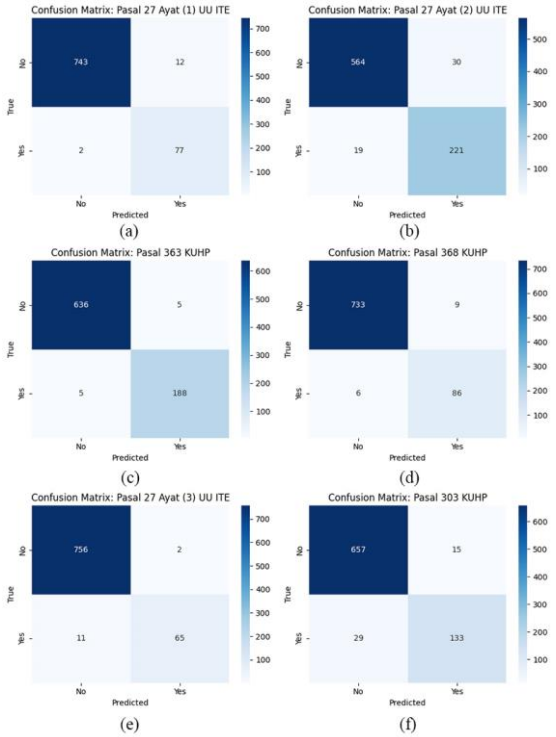

Figure 2. Confusion Matrix of IndoBert-base-p2

Commented [LA3]: Tambahkan (a), (b), (c),dst di bawah gambar untuk gambar lebih dari 1 dalam 1 caption

A confusion matrix was used to examine how well the models predicted each class by comparing actual labels to predicted ones. Looking at Figure 2(a), for Pasal 27 Ayat (1) UU ITE, the model correctly identified 77 true positives with only 2 false negatives, showing strong sensitivity for this article. However, the model incorrectly labeled 12 cases from other articles as Pasal 27 Ayat (1), possibly due to overlapping or similar language across legal descriptions.

Figure 2(b) shows that for Pasal 27 Ayat (2) UU ITE, the model struggled more, with 19 false negatives and 30 false positives. This higher error rate may result from subtle differences in meaning or a lack of distinctive features in the data for this article. This suggests a need for focused feature engineering or data augmentation to better capture the specific nuances of Pasal 27 Ayat (2).

Based on the figure 2 (e), for Pasal 27 Ayat (3) UU ITE, the confusion matrix shows a strong true positive rate (65) and relatively low false positives (2) and false negatives (11), indicating the model can distinguish this article well, though there remains some room for improvement in reducing false negatives.

Based on the figure 2 (f), Pasal 303 KUHP, the model detects a good number of true positives (133) but also shows notable false negatives (29) and false positives (15). This suggests that while the model is fairly accurate, it occasionally fails to recognize all relevant cases or mislabels others, possibly due to overlapping legal concepts or language patterns shared with other articles.

Based on the figure 2 (c), the classification for Pasal 363 KUHP performs very well, with 188 true positives and minimal false positives (5) and false negatives (5). This balanced confusion matrix reflects the model's robust ability to identify this article's cases reliably.

Based on the figure 2 (d), for Pasal 368 KUHP, the model yields 86 true positives, with 6 false negatives and 9 false positives. This again shows solid classification but indicates a slight tendency to confuse some cases with other classes, which could be addressed by refining model sensitivity or enriching training data.

In summary, the confusion matrices demonstrate that while the model achieves strong classification performance across most legal articles, certain classes, particularly Pasal 27 Ayat (2) UU ITE and Pasal 303

KUHP, show higher misclassification rates. These errors likely stem from semantic overlaps and data imbalances, highlighting opportunities for further model optimization.

D. Website Implementation

In this study, the best model of Bi-LSTM and IndoBERT is deployed as a web application using Streamlit, a Python-based framework designed for rapid and user-friendly interface creation. This implementation allows users to input case descriptions through a simple and interactive interface, receiving real-time predictions of relevant legal articles. The application is deployed locally, meaning it runs on a personal computer or local server rather than being hosted online, which simplifies testing and development without requiring web hosting infrastructure. Streamlit's flexibility and ease of integration with Python machine learning libraries enable seamless deployment of the trained models, making the legal classification tool accessible for demonstration and evaluation purposes. This web-based solution supports the study's goal of enhancing public understanding and accessibility of Indonesian legal information**.**

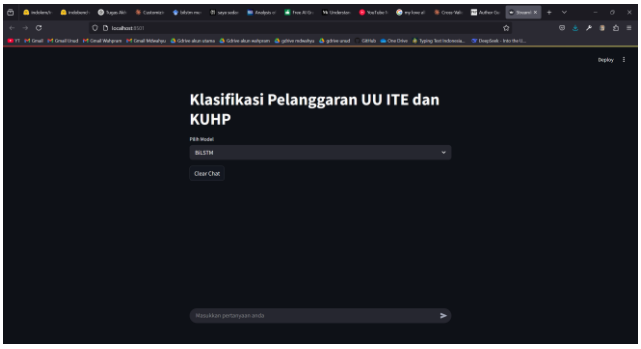

Figure 3. Landing Page Website

Figure 3 is the initial screen of the website prominently displays the website's title at the top, providing clear identification of the application. Below the title, users are presented with options to select the preferred model for legal article classification, enabling flexibility in model choice. A "Clear Chat" button is also available, allowing users to reset the input field and start a new query easily. The main interaction area features a text input bar, where users can enter case descriptions or legal questions to receive relevant legal article predictions.

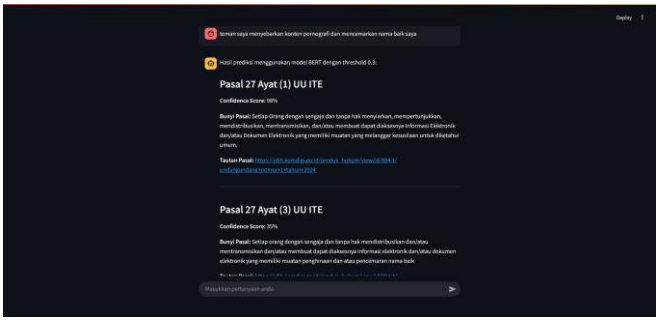

Figure 4. Model Prediction

Figure 4 is the image that illustrates an example interaction where a user inputs a legal case description into the text bar. The system responds with predicted relevant legal articles, each accompanied by a confidence score indicating the model's certainty in the prediction. Additionally, the full text of the corresponding legal article is displayed to provide context and clarity. These responses are presented in a chat bubble format, mirroring conversational interfaces, which enhances readability and user engagement by visually separating each predicted article and its details in an organized manner.

E. System Testing

Black-box testing was carried out to assess the functionality of the system by evaluating whether each feature operated according to its intended requirements. This testing focused solely on the outputs generated in response to specific inputs without examining the internal code or structure, as the system was deployed only in a local environment. The results of the testing are presented in Table 7.

TABEL 7
BLACK-BOX TESTING RESULT

Test Case ID	Feature Tested	Input	Expected Output	Result
TC-01	Text Input Validation	Empty text input	System displays error message "Input cannot be empty" and prevents submission	Pass
TC-02	Text Input Validation	Valid text (e.g., "Seseorang menyebarkan berita bohong di media sosial")	System accepts input and proceeds to prediction	Pass
TC-03	Model Selection	User selects *BiLSTM* model	Prediction result is generated using BiLSTM model	Pass
TC-04	Model Selection	User selects *IndoBERT* model	Prediction result is generated using IndoBERT model	Pass
TC-05	Clear Cache Feature	User clicks "Clear Cache"	All cached prediction results are cleared; system returns to initial state	Pass
TC-06	Prediction Result	Input: "Seseorang menyebarkan berita bohong di media sosial"	System predicts relevant article(s) (e.g., KUHP Article 303)	Pass
TC-07	Prediction Explanation	After prediction	System displays explanation of why the article was selected	Pass
TC-08	Legal Article Link	After prediction	System provides hyperlink to full legal article content	Pass

Black-box testing was conducted to evaluate the core functionalities of the legal violation classification system. The testing covered key features including text input validation, model selection, cache clearing, prediction generation, explanation display, and access to legal article links. Each test case was executed by providing specific inputs and verifying whether the outputs matched the expected behavior. As shown in Table 7, all eight test cases successfully passed, indicating that the system performed as intended.

IV. Conclusion

This study successfully developed a multilabel classification model to identify relevant legal articles from case descriptions based on the Indonesian Penal Code (KUHP) and the Information and Electronic Transactions Law (UU ITE). Two models were explored: Bidirectional Long Short-Term Memory (Bi-LSTM) and IndoBERT. Evaluation results showed that IndoBERT outperformed Bi-LSTM in terms of accuracy, F1-score, and hamming loss. However, the difference in F1-score was not statistically significant based on the t-test, indicating that Bi-LSTM remains a competitive model, especially in certain contexts.

Cross-validation results reinforced model consistency, with IndoBERT achieving an average accuracy of 97% and a hamming loss of 0.027, while Bi-LSTM achieved 90% accuracy and a hamming loss of 0.029. The confusion matrix analysis revealed that the model was generally reliable in classifying articles with balanced data but still experienced misclassifications, particularly in articles that are semantically similar.

However, this study is constrained by several limitations, including class imbalance and the dominance of certain legal articles within the dataset, which may affect the model's ability to generalize to underrepresented articles. The website prototype developed using Streamlit demonstrated that the model could be accessed through a simple and responsive interface, although the system is currently deployed only locally.

For future work, several directions are proposed: expanding the scope of legal articles to cover a broader range of cases, adapting the model to legal-specific BERT variants such as Legal-ID BERT for improved contextual understanding, and integrating calibration techniques as well as explainability methods (e.g., LIME or SHAP) to enhance system accountability in legal decision support.

References

[1] A. Perdana Hesaputra and D. Hatta Fudholi, "Klasifikasi Pelanggaran Undang-Undang ITE pada Twitter Menggunakan LSTM dan BiLSTM." [Online]. Available: https://t.co/0dnpcgQiF9

[2] M. Dhafa Maulana, C. Sri, and K. Aditya, "Perbandingan IndoBERT dan Bi-LSTM Dalam Mendeteksi Pelanggaran Undang-Undang ITE," *SINTECH JOURNAL*, vol. 8, no. 1, pp. 52–59, 2025, [Online]. Available: https://doi.org/10.31598

[3] A. D. Hasyim and Darsinah, "The Urgency of the Second Amendment to ITE Law from the Standpoint of the Positive Law and Human Rights," *Samarah*, vol. 9, no. 1, pp. 45–62, Mar. 2025, doi: 10.22373/sjhk.v9i1.22656.

[4] R. Hayami, "Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning Dan Deep Learning: Studi Literatur," 2023. [Online]. Available: https://ieeexplore.ieee.org/

[5] J. Amalia, J. Pakpahan, M. Pakpahan, Y. Panjaitan, F. Informatika dan Teknik Elektro, and I. Teknologi Del, "Model Klasifikasi Berita Palsu Menggunakan Bidirectional LSTM Dan Word2Vec Sebagai Vektorisasi," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 4, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[6] E. Aurora, A. Zahra, Y. Sibaroni, & Sri, and S. Prasetyowati, "Classification of Multi-Label of Hate Speech on Twitter Indonesia using LSTM and BiLSTM Method," *JINAV: Journal of Information and Visualization*, vol. 4, no. 2, pp. 2746–1440, 2023, doi: 10.35877/454RI.jinav1864.

[7] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," in *Procedia Computer Science*, 2022. doi: 10.1016/j.procs.2022.12.188.

[8] F. Farhan, T. Triase, and A. M. Harahap, "Penggunaan Algoritma Naive Bayes Dalam Text Mining Untuk Klasifikasi Pasal UU ITE," *J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD)*, vol. 6, no. 2, 2023, doi: 10.53513/jsk.v6i2.7896.

[9] J. Chen, "BiLSTM-enhanced legal text extraction model using fuzzy logic and metaphor recognition," *PeerJ Comput Sci*, vol. 11, 2025, doi: 10.7717/peerj-cs.2697.

[10] A. Rozaq *et al.*, "Legal Literacy in Indonesia: Leveraging Semantic-Based AI and NLP for Enhanced Civil Law Access," in *E3S Web of Conferences*, EDP Sciences, Apr. 2025. doi: 10.1051/e3sconf/202562203002.

[11] A. Fabian Azmi, A. Voutama, S. Karawang Jl HSRonggo Waluyo, and T. Timur, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix," vol. 13, no. 1, 2024.

[12] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - Journal of Business and Audit Information Systems*, vol. 4, no. 2, Aug. 2021, doi: 10.30813/jbase.v4i2.3000.

[13] N. Pandey, P. K. Patnaik, and S. Gupta, "Data Pre Processing for Machine Learning Models using Python Libraries," *Int J Eng Adv Technol*, vol. 9, no. 4, pp. 1995–1999, Apr. 2020, doi: 10.35940/ijeat.D9057.049420.

[14] D. R. Alghifari, M. Edi, and L. Firmansyah, "Implementasi Bidirectional LSTM untuk Analisis Sentimen Terhadap Layanan Grab Indonesia," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 2, pp. 89–99, Sep. 2022, doi: 10.34010/jamika.v12i2.7764.

[15] A. F. Al Farizi and Y. Sibaroni, "Implementation of BiLSTM and IndoBERT for Sentiment Analysis of TikTok Reviews," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 96–106, Jan. 2025, doi: 10.29100/jipi.v10i1.5815.

[16] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 2016. doi: 10.1109/IACC.2016.25.