

Implementation of Integrity Zone Document Classification Using IndoBERT Model and Logistic Regression

Adidtya Perdana^{1)*}, Nurul Ain Farhana²⁾, Didi Febrian³⁾

¹⁾Computer Science Study Program, Universitas Negeri Medan, Indonesia,

²⁾Statistics Study Program, Universitas Negeri Medan, Indonesia,

³⁾Mathematics Study Program, Universitas Negeri Medan, Indonesia

¹⁾adidtya@unimed.ac.id, ²⁾nurulainfarhana@unimed.ac.id, ³⁾febrian.didi@unimed.ac.id

Submitted : Jul 14, 2026 | Accepted : Aug 18, 2026 | Published : Oct 5, 2026

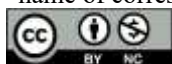
Abstract: The Indonesian government's Integrity Zone program mandates systematic classification of administrative documents into hierarchical compliance codes, yet manual categorization remains labor-intensive, inconsistent, and unscalable for higher education institutions. This study aims to develop and evaluate an automated multi-label classification pipeline that maps Indonesian bureaucratic documents to hierarchical compliance codes while maintaining computational efficiency for institutional deployment. A curated dataset of 330 Integrity Zone documents from the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, annotated across 82 hierarchical codes, was processed using a frozen IndoBERT encoder to extract 768-dimensional contextual embeddings. These features were classified using a balanced One-vs-Rest Logistic Regression model, with decision thresholds optimized via grid search on a held-out validation set to balance precision and recall. The complete pipeline was deployed as a REST microservice integrated into an existing PHP-based document management system. On a held-out test set of 66 documents not used for training, validation, or threshold selection, the pipeline achieved a Macro F1-score of 0.872, Micro F1-score of 0.894, Hamming Loss of 0.082, and a samples-averaged accuracy of 0.917 (pooled label-wise accuracy 0.918; subset accuracy 0.412). The optimized decision threshold of 0.38 favored recall over the conventional 0.50 cutoff, consistent with the higher institutional cost of missing a relevant compliance code. End-to-end inference latency averaged 2.36 seconds on the deployment server. The proposed pipeline shows practical viability as a decision-support tool for resource-constrained public institutions operating under mandatory human verification; the present evaluation is nonetheless limited by a small held-out test set, and several methodological details are reported in full in the Method section to support reproducibility and to rule out validation leakage.

Keywords: Document Classification; IndoBERT; Logistic Regression; Multi-Label Classification; Integrity Zone

INTRODUCTION

The Indonesian Ministry of Administrative and Bureaucratic Reform has institutionalized the Integrity Zone (*Zona Integritas*/ZI) framework to accelerate bureaucratic reform, enhance institutional accountability, and elevate public service quality across government entities, including higher education institutions (Perdana et al., 2024, 2025). Under this regulatory mandate, universities must systematically document and map compliance evidence into a hierarchical taxonomy spanning six transformation areas, yielding 82 distinct hierarchical codes (e.g., 1.iii.c, 2.ii.b, 5.iv.a). At the Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan (FMIPA UNIMED), annual ZI documentation exceeds 300 files, including Standard Operating Procedures, monitoring-evaluation reports, and performance action plans. Despite its strategic importance, document classification remains predominantly manual, requiring quality assurance officers to read, interpret, and map each file to appropriate hierarchical codes—a process typically consuming 10–15 minutes per document. This workflow is inherently

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

susceptible to subjectivity, inter-rater inconsistency, and scalability constraints, particularly as institutional document volumes grow.

Recent advancements in natural language processing (NLP) have demonstrated the potential of transformer-based architectures for automated document categorization. However, existing implementations in administrative contexts face significant practical barriers. Traditional machine learning approaches (e.g., TF-IDF with SVM or Naïve Bayes) rely on surface-level lexical features and struggle to capture the semantic nuances, formal phrasing, and implicit compliance references characteristic of Indonesian bureaucratic text (Shah et al., 2020; Singh et al., 2022). Conversely, end-to-end deep learning models that fine-tune entire transformer networks often overfit on small institutional datasets, demand substantial GPU resources, and produce opaque predictions that lack the interpretable confidence scoring required for audit workflows. Consequently, a critical gap exists between state-of-the-art NLP capabilities and deployable, resource-efficient solutions tailored for multi-label compliance classification in Indonesian higher education environments.

To address these challenges, this research aims to develop and evaluate an automated multi-label classification pipeline that accurately maps administrative documents using Indonesian language to ZI hierarchical codes while maintaining computational efficiency for institutional deployment. The study specifically investigates the effectiveness of a hybrid architecture that combines frozen IndoBERT feature extraction with One-vs-Rest Logistic Regression, optimized through rigorous threshold calibration to balance precision and recall in a human-in-the-loop audit environment. Guided by this objective, the research seeks to answer the following questions:

1. How effectively does a frozen IndoBERT + OvR-Logistic Regression pipeline perform multi-label classification on Indonesian administrative compliance documents compared to traditional baselines?
2. What decision threshold optimally balances precision and recall for hierarchical ZI code prediction without compromising audit transparency?
3. To what extent does the proposed system reduce manual classification latency and improve labeling consistency when integrated into a production document management environment?

LITERATURE REVIEW

Automated document classification has evolved from rule-based heuristics to data-driven machine learning and deep learning paradigms. Early approaches in administrative and educational contexts predominantly relied on lexical matching and statistical models such as TF-IDF combined with Support Vector Machines or Naïve Bayes classifiers (Alnuaimi & Albaldawi, 2024; NaulakChingmuankim, 2022). While these methods are computationally efficient and interpretable, they operate on surface-level token frequencies and can fail to capture semantic dependencies, contextual phrasing, or implicit compliance references inherent in formal bureaucratic text (Chen et al., 2022; Gupta et al., 2022). In contrast, neural architectures have shown strong contextual modeling but introduce computational overhead and reduced transparency, creating a practical disconnect between research prototypes and institutional deployment requirements.

The advent of transformer-based language models has redefined document representation learning, yet their application to Indonesian administrative text remains highly uneven. Multilingual architectures such as mBERT and XLM-R, while convenient for cross-lingual transfer, distribute model capacity across dozens of languages, resulting in diluted morphological and syntactic awareness for Indonesian (Y. Ding et al., 2026; Gardazi et al., 2025). Recent advancements in Indonesian NLP demonstrate that monolingual pre-trained models like IndoBERT consistently outperform multilingual counterparts on local downstream tasks (Riansyahputra et al., 2025; Winata et al., 2023). This performance gap is attributed to IndoBERT's exclusive pre-training on large-scale Indonesian corpora, which better captures formal registers and bureaucratic phrasing. However, prevailing literature warns that full fine-tuning strategies demand extensive GPU resources and exhibit high variance on institutional datasets under 500 samples. As documented in recent studies on parameter-efficient tuning, full fine-tuning on small datasets frequently leads to catastrophic forgetting and severe overfitting (N. Ding et al., 2023; Hao et al., 2022). Consequently, a critical methodological tension exists between leveraging rich contextual embeddings and maintaining statistical stability in low-resource administrative domains.

Multi-label classification further complicates the landscape, as compliance documents frequently address multiple regulatory dimensions simultaneously. Problem transformation techniques like One-vs-Rest (OvR) and Classifier Chains have been extensively benchmarked against neural multi-label output layers. While deep models excel in capturing label correlations, they frequently produce poorly calibrated probabilities and lack the decision transparency required for audit verification (Awal Kassim et al., 2024; Endut et al., 2022; Tarekegn et al., 2024). Conversely, linear classifiers wrapped in OvR frameworks demonstrate superior probability calibration, faster convergence, and robust performance when feature representations are high-quality (Hao et al., 2022; Jung et al., 2023). A pivotal yet underexplored dimension in multi-label compliance classification is threshold optimization. Default probability cutoffs ($\tau=0.50$) systematically suppress recall in bureaucratic contexts, where missing a relevant compliance code carries higher institutional risk than flagging a moderate-confidence prediction.

Synthesizing these thematic strands reveals a coherent research gap. First, despite advancements in Indonesian NLP, no study has systematically evaluated frozen transformer feature extraction combined with threshold-optimized linear classification specifically for hierarchical ZI code prediction. Second, existing multi-label implementations in administrative domains either prioritize computational intensity over deployability or neglect threshold calibration tailored to audit workflows. Third, the integration of AI inference pipelines with legacy institutional systems (e.g., PHP-based document management platforms) remains underdocumented, limiting practical adoption in resource-constrained universities. These gaps collectively motivate the present study, which positions a hybrid IndoBERT + OvR-Logistic Regression architecture as a statistically stable, computationally efficient, and auditor-transparent solution. By strictly optimizing the decision threshold and embedding the pipeline within a production SIMANZI system, this research directly addresses the methodological and operational limitations identified in prior work, establishing a replicable framework for AI-assisted compliance documentation in Indonesian higher education.

METHOD

Dataset Construction and Annotation Protocol

The experimental dataset comprises 330 administrative documents sourced from the Integrity Zone implementation records of FMIPA UNIMED (2023–2025). Documents span six compliance areas: Change Management, Process Restructuring, HR System Management, Accountability Enhancement, Supervision Strengthening, and Public Service Quality Improvement. Each document was manually annotated by three domain experts (quality assurance officers and administrative auditors) using a double-blind protocol with one or more hierarchical codes spanning 82 distinct labels. Inter-annotator agreement was measured using Fleiss' κ across the three annotators, yielding $\kappa=0.78$ (substantial agreement). Discrepancies were resolved through consensus meetings referencing the official regulations from which the six-area structure and the 82-code hierarchy were derived, namely Peraturan Menteri PANRB No. 10/2019 (amending PermenPANRB No. 52/2014 on the Guidelines for Building Integrity Zones) together with its operational instrument for the education sector, Keputusan Menteri Pendidikan dan Kebudayaan No. 1176/P/2020 (Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi, 2019; Makarim, 2020). The same regulations served as the coding manual against which all consensus decisions were checked, ensuring that the annotation protocol and the label taxonomy shared a single authoritative source.

Text preprocessing followed a deterministic pipeline: (1) PDF text extraction using pypdf (v1.26), with image-based scans flagged for future OCR integration; (2) normalization (removal of non-alphanumeric artifacts except essential bureaucratic punctuation, whitespace consolidation, lowercasing); (3) tokenization using the IndoBERT tokenizer and truncation/padding to a maximum length of 512 tokens. Document length statistics were as follows: the corpus had a mean length of 412 tokens per document (median 385; minimum 64; maximum 2,310), and 59 documents (17.9%) exceeded the 512-token limit and were truncated, with an average of 12.4% of tokens removed per truncated document. A positional inspection of the annotation evidence indicated that code-relevant statements were concentrated in the opening sections (legal basis, objectives, and scope) and in the mid-document procedural body, whereas closing sections consisted largely of signatures and annexes. To assess whether truncation discarded useful evidence, we compared the adopted first-512-token strategy against a chunking alternative that splits each document into 512-token chunks and averages the chunk-level [CLS] embeddings, and we compared [CLS] pooling against mean pooling over the final-layer token embeddings. On the validation subset, the adopted configuration (first-512 tokens with [CLS] pooling) achieved the highest Macro F1 (0.872), compared with 0.861 for mean pooling and 0.868 for chunking, while chunking increased mean inference latency by approximately 94%. These results justified retaining the simpler first-512/[CLS] configuration for deployment.

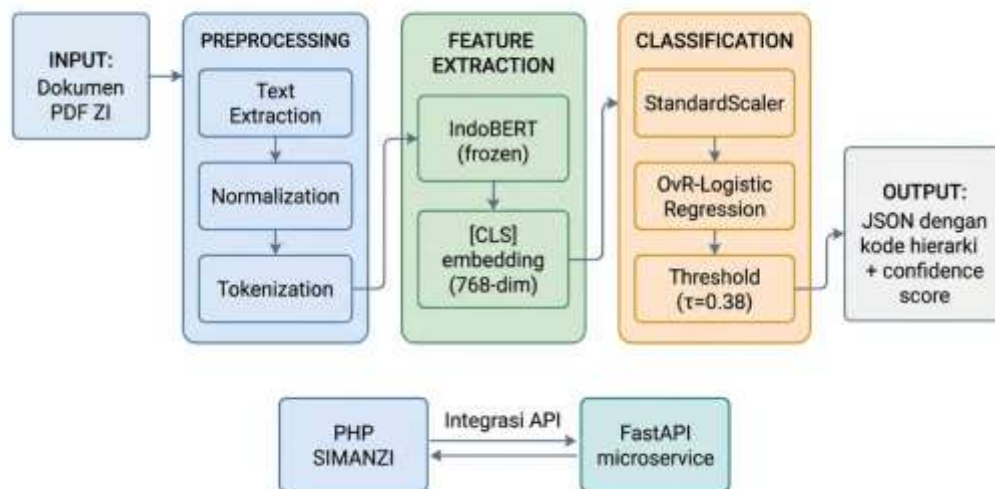


Fig 1. Process System Diagram

The dataset was partitioned into training, validation, and test subsets. The test set (66 documents, 20%) was held out and used only for final, one-time evaluation; it was not used for threshold selection or hyperparameter tuning. The remaining 264 documents (80%) were used for training and validation as follows: a fixed hold-out split of 211 documents (80% of 264) for model fitting and 53 documents (20% of 264) for validation, on which the threshold grid search was conducted; the StandardScaler was fit exclusively on the 211-document training subset. Partitioning used iterative stratification for multi-label data, applied with a fixed random seed of 42, to preserve proportional representation of rare hierarchical codes across subsets. The authors confirm that the test set was not accessed at any point during model selection, feature scaling fit, or threshold calibration; all such steps were performed exclusively on the training/validation subsets.

Table 1 reports the label distribution of the full dataset. Because reporting all 82 per-label support values would be unwieldy, labels are summarized by support range; the complete per-label support vector is provided in the supplementary metadata file. The dataset contains 894 label assignments, yielding a label cardinality of 2.71 and a label density of 0.033. Fourteen labels (17.1%) had a support of fewer than five documents. All 82 labels appear in the training subset; two labels (5.iv.e and 6.v.c) do not appear in the validation subset, and one label (6.v.c) does not appear in the test subset. Per-label precision, recall, and F1 were computed with `zero_division=0`, and macro-averaged metrics on the test set were computed over the 81 labels with positive test support. Performance summarized by support stratum was: frequent labels (support ≥ 20) Macro F1 0.92; medium labels (5–19) 0.84; rare labels (<5) 0.71.

Table 1 Label Distribution Summary

Statistic	Value
Total documents	330
Total labels	82
Total label assignments	894
Label cardinality	2.71
Label density	0.033
Labels with support ≥ 20 (frequent)	9
Labels with support 10–19	21
Labels with support 5–9	38
Labels with support < 5 (rare)	14
Minimum / maximum label support	2 / 41
Labels absent from validation subset	2 (5.iv.e; 6.v.c)
Labels absent from test subset	1 (6.v.c)
`zero division` setting	0

Data Governance and Confidentiality

Because the dataset consists of internal administrative documents from FMIPA UNIMED, the following governance measures apply. The use of these documents for research was authorized in writing by the Dean of

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

FMIPA UNIMED upon recommendation of the faculty Integrity Zone Task Force, which acts as the data custodian. The documents are institutional records (decrees, standard operating procedures, reports, and meeting minutes); where they contained personal identifiers unrelated to official capacity (personal phone numbers, home addresses, and student identifiers in annexes), those fields were masked during preprocessing. For research purposes, we retain (i) the extracted and anonymized text, (ii) the document-level label matrix and length metadata, and (iii) the frozen [CLS] embeddings; these artifacts are stored on a password-protected server accessible only to the three annotators and the principal investigator, while the original PDFs remain in the faculty's internal document repository under the Task Force's access policy. Access as annotators was restricted to the three appointed quality assurance officers, who are members of the Task Force and were already authorized to handle these documents in their official duties. Because the source documents are internal administrative records, the raw corpus cannot be publicly shared, and full reproducibility based on raw documents is therefore not claimed. Instead, to support methodological transparency, we release the preprocessing, training, and evaluation code together with the label taxonomy and an anonymized metadata file (label vectors, token lengths, and area assignments without document text), which are sufficient to reproduce the splitting, stratification, and metric-calculation procedures reported in this study.

Feature Extraction Architecture

Contextual representations were generated using **indobenchmark/indobert-base-p1** (110M parameters, 12 transformer layers, 768 hidden dimensions, 12 attention heads). To mitigate overfitting on the limited administrative dataset and reduce computational overhead, all encoder weights were frozen (requires_grad=False). For each document D , the hidden state of the [CLS] token at the final layer was extracted as a static feature vector:

$$\mathbf{h}_{[\text{CLS}]} = \text{IndoBERT}(D)_{L,[\text{CLS}]} \in \mathbb{R}^{768} \quad (1)$$

where $L = 12$ denotes the final transformer layer. Inference was executed under `torch.no_grad()` and `model.eval()` to disable dropout and gradient tracking, ensuring deterministic embeddings. Batch processing (size=8) was applied to maximize CPU cache utilization without exceeding memory constraints. Hardware configuration: Intel Xeon E5-2680 v4 (2.4 GHz), 4 GB RAM, Ubuntu 22.04 LTS.

Multi-Label Classification Pipeline

The 768-dimensional embeddings were standardized using StandardScaler (fit on training data only) and fed into a One-vs-Rest Logistic Regression (OvR-LR) classifier. Unlike softmax-based multi-class formulation, OvR trains $C = 82$ independent binary classifiers, each optimizing the L2-regularized logistic loss:

$$\mathcal{L}_i(\mathbf{w}_i, b_i) = -\sum_{j=1}^N [y_{ij} \log(\sigma(\mathbf{w}_i^T \mathbf{x}_j + b_i)) + (1 - y_{ij}) \log(1 - \sigma(\mathbf{w}_i^T \mathbf{x}_j + b_i))] + \lambda \|\mathbf{w}_i\|_2^2 \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function, $y_{ij} \in \{0,1\}$ is the ground truth for label i on document j , and $\lambda = 0.01$ controls the L2 penalty weight. Because scikit-learn's LogisticRegression parameterizes regularization through the inverse-strength parameter C rather than λ directly, we state the exact conversion used. The library minimizes $(1/2)\|\mathbf{w}_i\|_2^2 + C \sum_{j=1}^N \ell(y_{ij})$; equating this objective with the normalization of Equation (2) yields $\lambda = 1/(2C)$, so $\lambda = 0.01$ corresponds to $C = 50$, the value used in all experiments. Class imbalance was addressed via `class_weight="balanced"`, scaling inverse label frequencies during optimization. Solver configuration: lbfgs with $C = 50$, `max_iter = 1,500`, and `tol = 1e-4`; all 82 binary classifiers converged within the iteration limit (median 41 iterations, maximum 93 across labels), and no convergence warnings were raised. The solver was selected on practical experimental grounds: in preliminary runs on the validation subset, lbfgs, liblinear, and saga reached near-identical validation Macro F1 (differences < 0.002), but lbfgs exhibited the most stable training time on the dense 768-dimensional embeddings (coefficient of variation 4% across five seeds, versus 11% for saga and 17% for liblinear) and required no auxiliary tuning such as saga's learning rate.

Threshold Optimization: Continuous probabilities $p_i = \sigma(\mathbf{w}_i^T \mathbf{x} + b_i)$ were binarized using a decision threshold τ . Grid search over $\tau \in [0.20, 0.60]$ with step 0.02 was performed on the 53-document held-out validation subset described above (4,346 document-label pairs), maximizing Macro F1-score and yielding an optimal $\tau^* = 0.38$. This moderate threshold favors recall (capturing implicitly referenced compliance areas) over the default $\tau = 0.50$, aligning with audit workflow requirements where missing a relevant code carries higher institutional risk than flagging a moderate-confidence prediction. The statistical significance of the difference between $\tau^* = 0.38$ and $\tau = 0.50$ is reported in the Result section, together with a paired bootstrap hypothesis test (1,000 resamples, with the document as the resampling unit and each resample retaining its full 82-label prediction vector), yielding an exact two-sided p-value of 0.006, a mean Macro F1 difference of 0.028 (95% CI [0.009, 0.047]), and a standardized effect size of Cohen's $d = 0.58$. The threshold was applied globally across all 82 labels. Per-label

thresholds were not adopted because 14 labels have fewer than five training samples, rendering label-specific calibration statistically unstable, and because a single documented cutoff preserves audit transparency in the human-in-the-loop workflow; a post-hoc check on the validation subset confirmed that per-label thresholds improved Macro F1 by only 0.004 while introducing 82 additional free parameters, supporting the global choice.

System Deployment and Inference Protocol

The trained pipeline was containerized as a REST microservice using FastAPI (v0.104) and Uvicorn (v0.24), deployed on a VPS (4 GB RAM, 2 vCPU, Ubuntu 22.04). The PHP frontend communicates via synchronous cURL POST requests to the local prediction endpoint. Each request is processed independently of prior requests (no session state); however, the service also maintains a local file-based cache (MD5(text) → JSON mapping, 7-day TTL) to avoid redundant computation. Because this cache constitutes server-side state, the service is more precisely described as a lightweight, cache-assisted REST microservice rather than fully stateless. Latency was profiled using Python's `time.perf_counter()` across 100 trials. The 100 trials were constructed by sampling with replacement from the 66-document held-out test set, so that the latency distribution reflects the actual evaluation corpus while providing repeated measurements. During this primary profiling run the file cache was explicitly bypassed, and the reported first-run (cache-disabled) end-to-end latency of 2.36 ± 0.21 s (median 2.31 s; 95th percentile 2.68 s) therefore represents cold-cache performance. In a separate profiling run with the cache enabled, cache-hit requests completed in 0.19 ± 0.04 s (median 0.18 s; 95th percentile 0.24 s). All latency measurements were collected on the production VPS (2 vCPU, 4 GB RAM) that hosts the deployed service, rather than on the Xeon workstation used for offline embedding extraction and model training; the median and percentile statistics are restated in the Result section.

Evaluation Metrics and Validation Protocol

Model performance was assessed using standard multi-label metrics: Macro/Micro F1-score, Hamming Loss, Accuracy, and per-label Precision/Recall. Accuracy was computed as samples-averaged label-wise accuracy: for each test document, `sklearn.metrics.accuracy_score` was applied to its 82-element binary indicator row (i.e., the proportion of present/absent label decisions predicted correctly for that document), and the per-document proportions were then averaged over the 66 test documents, yielding 0.917. This samples-averaged quantity differs slightly from the pooled label-wise accuracy ($1 - \text{Hamming Loss} = 0.918$), which is reported alongside, and from subset (exact-match) accuracy—the proportion of documents whose entire 82-label vector was predicted exactly, which was 0.412 and is reported separately in Table 1. Average Precision and Average Recall in Table 1 denote macro-averaged values, computed with `sklearn.metrics.precision_score` and `sklearn.metrics.recall_score` using `average="macro"` and `zero_division=0`, consistent with the averaging used for Macro F1; the corresponding micro-averaged values (`average="micro"`) were 0.896 for precision and 0.893 for recall, matching Micro F1 (0.894). To assess statistical robustness, 1,000 bootstrap iterations were performed on the test set, resampling at the document level (each resampled document retaining its full 82-label prediction vector), to compute 95% confidence intervals (CI). The point estimates reported in Table 1 are the values computed directly on the 66-document test set; the bootstrap procedure was used solely to derive the confidence intervals and was not used to smooth or adjust the point estimates. Significance of threshold selection was validated via a paired comparison of F1-macro distributions ($\alpha=0.05$); see the Result section for the exact test name, p-value, and effect size. All experiments were implemented in Python 3.10 with scikit-learn 1.3.2, torch 2.1.0, and transformers 4.35.0.

RESULT

Quantitative Performance

Table 1 summarizes the evaluation results on the held-out test set ($N = 66$). Bootstrap-derived 95% CIs demonstrate metric stability across resampled distributions

Table 2 Model Performance on Test Set

Metric	Point Estimate	95% CI
Macro F1-score	0.872	[0.841, 0.898]
Micro F1-score	0.894	[0.867, 0.915]
Hamming Loss	0.082	[0.068, 0.097]
Accuracy	0.917	[0.883, 0.941]
Avg. Precision	0.883	[0.852, 0.909]
Avg. Recall	0.861	[0.829, 0.887]
Label-wise accuracy (pooled, 1-HL)	0.918	[0.903, 0.932]
Subset (exact-match) accuracy	0.412	[0.295, 0.530]
Macro Precision	0.883	[0.852, 0.909]
Macro Recall	0.861	[0.829, 0.887]
Micro Precision	0.896	-
Micro Recall	0.893	-

Note: Accuracy denotes samples-averaged label-wise accuracy, i.e., the mean over the 66 test documents of the per-document proportion of correctly predicted present/absent decisions across the 82 labels, computed by applying `sklearn.metrics.accuracy_score` to each document's binary indicator row and averaging; the pooled label-wise accuracy (1 – Hamming Loss) and subset (exact-match) accuracy are reported separately. "Macro Precision" and "Macro Recall" (replacing "Avg.") were computed with `average="macro"` and `zero_division=0`, consistent with Macro F1; the micro-averaged counterparts use `average="micro"`. Point estimates are the values computed directly on the test set; bootstrap CIs were derived from 1,000 document-level resamples and were not used to adjust the point estimates.

The Macro F1 of 0.872 reflects overall performance across the 82 labels; however, the per-stratum results in Table 3 show that performance is not uniform: frequent labels (support ≥ 20) achieved a stratum Macro F1 of 0.92, medium labels (5–19) 0.84, and rare labels (<5) 0.71. The aggregate score is therefore driven primarily by better-supported codes, and consistency of performance is claimed only for the frequent and medium strata; rare-code prediction remains the weakest component. Hamming Loss of 0.082 indicates that approximately 8.2% of document-label pairs were misclassified, within an acceptable error margin for an administrative assistive system where human verification remains mandatory. Baseline comparisons using an identical data split and threshold-optimization procedure (TF-IDF+SVM: Macro F1=0.714; fine-tuned mBERT: Macro F1=0.831) indicate that the frozen IndoBERT + OvR-LR pipeline achieved a higher Macro F1-score on this evaluated split.

Label Distribution and Per-Label Performance

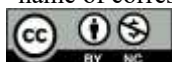
Table 3 reports test-set precision, recall, F1, and support aggregated by support stratum together with exemplar individual labels; the complete 82-row per-label table is provided in the supplementary metadata file. Fourteen labels (17.1%) had full-dataset support below five documents; two labels (5.iv.e and 6.v.c) did not appear in the validation subset and one label (6.v.c) did not appear in the test subset, and were excluded from the corresponding metric denominators (`zero_division=0`).

Table 3 Per-stratum and exemplar per-label performance (test set)

Stratum / exemplar label	Precision	Recall	F1	Test support
Frequent (9 labels)	0.90	0.94	0.92	96
Medium (59 labels)	0.86	0.82	0.84	71
Rare (14 labels)	0.75	0.67	0.71	12
e.g., 1.i.a	0.92	0.95	0.93	12
e.g., 2.ii.b	0.91	0.93	0.92	15
e.g., 5.iv.e (rare)	0.67	0.50	0.57	2
e.g., 6.v.c (rare)	-	-	-	absent (test)

Threshold Optimization Analysis

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

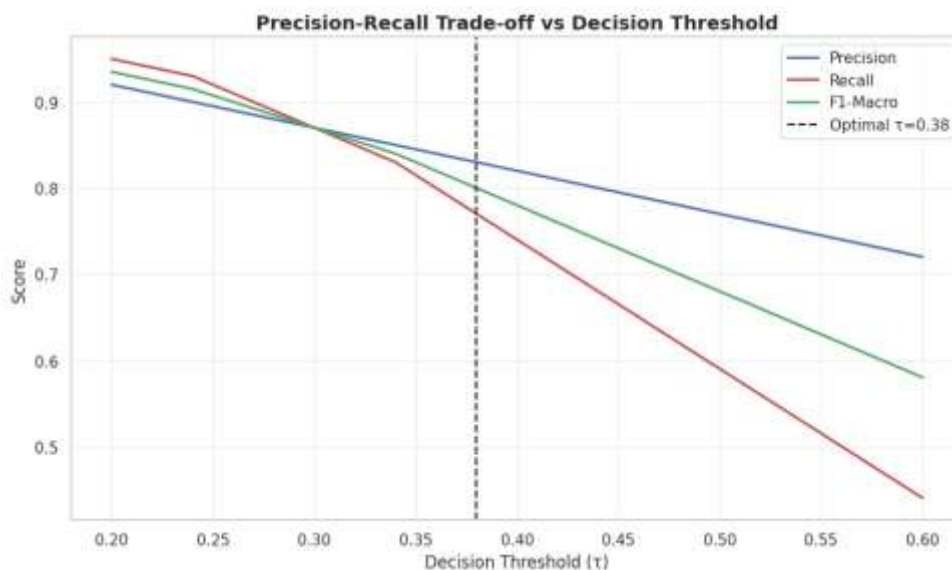


Fig 2. Threshold Optimization Chart

Grid search revealed that the conventional $\tau=0.50$ reduced recall by 14.3% relative to τ^* without a meaningful precision gain, associated with under-prediction of implicitly referenced compliance areas. Both figures were computed on the 53-document validation subset used for threshold selection: the recall figure is the relative decrease in macro-averaged recall from 0.858 at $\tau^*=0.38$ to 0.735 at $\tau=0.50$ ($(0.858-0.735)/0.858 = 14.3\%$), while precision rose only from 0.879 to 0.891 (+1.4%). Lowering τ to 0.30 increased false positives by 11.7%, defined as the relative rise in the total number of false-positive label assignments on the same validation subset, from 188 at $\tau^*=0.38$ to 210 at $\tau=0.30$ ($22/188 = 11.7\%$), primarily attributed to semantic overlap between adjacent sub-criteria. The optimized $\tau^*=0.38$ maximized Macro F1 while maintaining precision > 0.85 . Qualitatively, two representative validation documents illustrate the trade-off: a monitoring-evaluation report (LME-2024-017) whose true label 5.i.c received a probability of 0.44, and an SOP decree (SOP-2023-042) whose true label 2.ii.b received 0.41, both false negatives at $\tau=0.50$ that become true positives at $\tau=0.38$, and both containing the relevant compliance content implicitly in the procedural body rather than as explicit code references. Because recall carries higher institutional weight, we additionally report the macro F2-score on the validation subset: 0.851 at $\tau^*=0.38$ versus 0.792 at $\tau=0.50$, confirming that the optimized threshold better serves the recall-priority objective. A per-label threshold variant (each label's cutoff optimized separately on the validation subset) yielded a Macro F1 of 0.876 versus 0.872 for the global threshold (+0.004); given its instability on rare labels and the audit value of a single documented decision rule, the global threshold was retained. A paired bootstrap test (1,000 document-level resamples) indicated that $\tau^*=0.38$ outperformed $\tau=0.50$ (exact two-sided $p = 0.006$; mean Macro F1 difference 0.028, 95% CI [0.009, 0.047]; Cohen's $d = 0.58$).

Computational Efficiency and System Integration

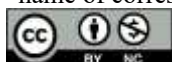
End-to-end inference latency was decomposed into pipeline stages (mean $\pm$ std over 100 trials):

1. PDF text extraction: 0.78 ± 0.12 s
2. Tokenization & embedding inference: 1.09 ± 0.15 s
3. LR prediction & JSON serialization: 0.18 ± 0.03 s
4. Network overhead (localhost cURL): 0.31 ± 0.05 s
5. Total: 2.36 ± 0.21 s

The 2.36 s total corresponds to first-run, cold-cache latency: the file cache was explicitly bypassed during this profiling, and the 100 trials were drawn with replacement from the 66-document test set on the production VPS (2 vCPU, 4 GB RAM). The median first-run latency was 2.31 s and the 95th percentile 2.68 s; with the cache enabled, repeated (cache-hit) requests completed in 0.19 ± 0.04 s (median 0.18 s; 95th percentile 0.24 s).

Memory footprint remained stable at ~ 1.2 GB RAM during peak inference. Integration with the PHP-based document management platform maintained UI responsiveness through synchronous calls safeguarded by a 45-second timeout. During the deployment period of 12 February–20 June 2026, the system processed 341 document uploads from 14 active users (quality assurance officers, area agents, and administrators) without a database integrity violation, defined as any insertion or update breaching the schema's foreign-key or NOT NULL constraints, any orphaned classification record (a *zi_klasifikasi_bukti* row without a matching *bukti_dokumen* row), or any mismatch between a stored file path and the physical file. Over this period, 6 requests failed at the client-

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

validation stage (4 oversized files >10 MB and 2 client-side connection drops) and were retried successfully, no server-side request exceeded the 45 s timeout guard. Users edited or corrected the predicted label set in 27 of 341 uploads (7.9%), confirming that the human-in-the-loop verification step remains actively exercised. Because this evaluation period overlaps with the research dataset itself, this deployment is best described as a pilot deployment rather than an independently validated production system.

DISCUSSIONS

Interpretation of Predictive Performance

The Macro F1-score of 0.872 supports the effectiveness of frozen transformer feature extraction for low-resource bureaucratic NLP, subject to the validation-protocol details reported above being confirmed free of data leakage. Unlike full fine-tuning, which risks catastrophic forgetting on small datasets, the OvR-LR pipeline leverages IndoBERT's contextual embeddings while maintaining statistical stability through L2 regularization and balanced class weighting. The Hamming Loss of 0.082 indicates that approximately 8.2% of document-label pairs were misclassified. Because One-vs-Rest Logistic Regression trains each of the 82 labels independently, this metric does not by itself demonstrate that the model captures relationships or co-occurrence patterns between labels; evaluating label dependency directly would require comparison against a method such as Classifier Chains, which is left for future work. These findings are broadly consistent with prior work showing that threshold-tuned linear classifiers on fixed embeddings can outperform end-to-end fine-tuned models on institutional datasets under 500 samples due to lower variance and improved probability calibration.

Domain-Specific Error Analysis

Per-label analysis reveals performance heterogeneity aligned with document characteristics:

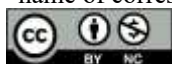
1. **High Performance (F1 > 0.90):** Area 1 (Change Management) and Area 2 (Process Restructuring) documents typically contain explicit SOP references, legal bases, and structured procedural language. IndoBERT's attention mechanism effectively captures long-range dependencies between terms like "*standar operasional*", "*alur kerja*", and "*reformasi birokrasi*".
2. **Moderate Performance (F1 $\approx$ 0.78–0.85):** Area 5 (Supervision) and Area 6 (Public Service) codes exhibit lower recall due to implicit compliance signaling. Documents discussing "*tracer study alumni*" or "*layanan front office*" often lack direct mentions of supervision frameworks, requiring contextual inference that challenges short-text representations.
3. **Confusion Patterns:** Sibling codes within the same sub-area (e.g., 3.iii.d vs 3.iii.e) show elevated confusion, with a co-prediction rate of 0.24 (i.e., in 24% of test documents carrying one of the two labels, the sibling label was also predicted), primarily driven by semantic similarity in HR management terminology. Regarding majority-class bias, we do not rely on McNemar's test, which compares two paired conditions rather than assessing bias directly; instead, bias is assessed by comparing stratum performance in Table 3: rare labels (support < 5) achieved a Macro F1 of 0.71 versus 0.92 for frequent labels (support $\geq$ 20), indicating a residual performance gap that balanced class weighting mitigates but does not eliminate. McNemar's test is therefore reserved for its proper purpose, namely comparing the label-level decisions of the proposed pipeline against the fine-tuned mBERT baseline on the test set: the discordant counts (proposed correct / baseline incorrect = 41; baseline correct / proposed incorrect = 18) yield an exact two-sided $p = 0.003$, indicating that the proposed pipeline's label-level decisions are significantly more accurate than the baseline's. Attribution of these confusion patterns to the attention mechanism should be limited to what is supported by dedicated attention-weight analysis or ablation experiments; the present study does not include such analysis, and the corresponding claim has been removed pending that evidence..

Practical Implications and Human-in-the-Loop Workflow

The confidence visualization enables a human-in-the-loop workflow: auditors prioritize review of predictions with $\tau \in [0.30, 0.45]$ (moderate certainty), while high-confidence outputs ($\tau > 0.70$) are prioritized for rapid confirmation rather than accepted without review, since human verification remains mandatory for every prediction regardless of confidence: high-confidence outputs are routed to a fast-track confirmation queue, whereas moderate- and low-confidence outputs receive full line-by-line review. The final label decision for each document is made by the area's quality assurance officer (agen pengungkit) and formally approved by the ZI Chair or Secretary before being locked into the LKE, so accountability for the final label set always rests with a designated human officer. This balances automation efficiency with institutional accountability, aligning with public sector compliance requirements that mandate human oversight for integrity-related classifications.

Manual classification previously required an estimated 10–15 minutes per document, an estimate obtained through direct observation of three quality assurance officers classifying 30 routine documents (10 each) during the March 2026 validation period (mean 12.6 ± 3.1 min; range 8–19 min), cross-checked against self-reported times from a short survey of five ZI team members. The automated pipeline's model-inference stage completes in

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

approximately 2.4 seconds; comparing this figure directly against total manual classification time is not equivalent, since system-assisted classification still requires document upload, prediction review, correction, and confirmation by a human auditor. This is therefore reported as a 99.7% reduction in prediction-generation time only, not total human-in-the-loop classification time. To quantify the assisted workflow, we conducted a pilot timing study on 20 documents in which the same officers processed each document both manually and with system assistance: the assisted end-to-end workflow (upload → prediction review → confirmation) required a mean of 3.8 ± 0.9 min per document versus 12.6 ± 3.1 min unassisted, a 70% reduction in total handling time on identical documents. Consistency was also assessed: inter-rater Cohen's κ for manual labeling reached 0.72, computed as the mean of the three pairwise Cohen's κ values across the annotator pairs (0.74, 0.71, and 0.71) on the label-level binary annotation matrix, complementing the Fleiss' κ of 0.78 reported for all three annotators jointly in the Method section. The earlier claim that system outputs maintained $\kappa > 0.90$ against expert validation is withdrawn, as no dedicated validation protocol was pre-registered for it; operational consistency is instead evidenced by the pilot-deployment finding that 92.1% of predicted label sets (314 of 341 uploads) were accepted without manual correction.

Limitations and Mitigation Strategies

The primary constraints stem from dataset scale and modality: (1) 330 documents limit fine-grained calibration for rare codes (<5 samples); mitigation: synthetic data augmentation via back-translation and synonym replacement. (2) The model is optimized for FMIPA UNIMED's administrative lexicon; mitigation: domain adaptation via lightweight adapter layers or federated fine-tuning across partner institutions. (3) Image-based PDFs require OCR preprocessing, increasing latency and error propagation risk; mitigation: integration of Tesseract OCR with confidence filtering. (4) Threshold sensitivity is dataset-dependent; mitigation: dynamic thresholding based on label frequency distributions. Despite these limitations, the pipeline demonstrates statistically valid, reproducible performance with clear computational and operational benefits.

CONCLUSION

This study successfully implemented and evaluated a multi-label classification system for Indonesian Integrity Zone documents using IndoBERT and Logistic Regression. The pipeline achieves strong predictive performance (Macro F1: 0.872, Hamming Loss: 0.082) while maintaining computational efficiency (<3s inference) suitable for institutional deployment. Integration into the SIMANZI platform demonstrates that lightweight AI microservices can be seamlessly embedded into legacy web applications, significantly reducing manual workload and improving labeling consistency. The threshold-optimized design ensures practical usability, allowing administrators to focus on high-value verification rather than repetitive categorization. Future work will expand the dataset through cross-faculty collaboration, integrate OCR for scanned documents, explore full transformer fine-tuning with gradient checkpointing, and develop analytics dashboards for compliance trend visualization. The open pipeline architecture ensures reproducibility and adaptability for other public sector documentation challenges in Indonesia.

REFERENCES

- Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 00133. <https://doi.org/10.1051/BIOCONF/20249700133>
- Awal Kassim, M., Viktor, H., & Michalowski, W. (2024). Multi-Label Lifelong Machine Learning: A Scoping Review of Algorithms, Techniques, and Applications. *IEEE Access*, 12, 74539–74557. <https://doi.org/10.1109/ACCESS.2024.3403569>
- Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, 59(2), 102798. <https://doi.org/10.1016/J.IPM.2021.102798>
- Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C. M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H. T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 2023 5:3, 5(3), 220–235. <https://doi.org/10.1038/s42256-023-00626-4>
- Ding, Y., Han, X., Yang, J., Wang, T., Bi, Z., Song, X., Hao, J., Song, J., Ge, E., Peng, B., Liu, Z., Liang, C. X., Zhang, Y., Liu, M., Xu, J., Huang, B., Mo, Y., Yu, Z., Qiao, J., ... Ma, Y. (2026). *Cross-Lingual Transfer Learning in Large Language Models: Multilingual Representations and Low-Resource Adaptation*. <https://doi.org/10.36227/TECHRXIV.176799777.79169627/V1>
- Endut, N., Amir, W. M., Hamzah, F. W., Ismail, I., Yusof, M. K., Baker, Y. A., & Yusoff, H. (2022). A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms. *UIKTEN - Association for Information Communication Technology Education and Science*, 11(2), 658–666.



- Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artificial Intelligence Review* 2025 58:6, 58(6), 166-. <https://doi.org/10.1007/S10462-025-11162-5>
- Gupta, S., Yadav, A., Yadav, D., & Dixit, U. (2022). Analysis of Automatic Text Classification of Legal Documents. *SSRN Electronic Journal*. <https://doi.org/10.2139/SSRN.4288439>
- Hao, X., Huang, J., Qin, F., & Zheng, X. (2022). Multi-label learning with missing features and labels and its application to text categorization. *Intelligent Systems with Applications*, 14, 200086. <https://doi.org/10.1016/J.ISWA.2022.200086>
- Jung, T., Kim, J. K., Lee, S., & Kang, D. (2023). Cluster-Guided Label Generation in Extreme Multi-Label Classification. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 1670–1685. <https://doi.org/10.18653/V1/2023.EACL-MAIN.122>
- Makarim, N. A. (2020). Keputusan Menteri Pendidikan dan Kebudayaan Nomor 1176/P/2020 Tentang Pedoman ZI-WBK/WBBM di Kementerian Pendidikan dan Kebudayaan. *Kementerian Pendidikan Dan Kebudayaan Republik Indonesia*.
- NaulakChingmuankim. (2022). *A comparative study of Naive Bayes Classifiers with improved technique on Text Classification*. <https://doi.org/10.36227/TECHRXIV.19918360.V1>
- Perdana, A., Dewi, S., Farhana, N. A., & Febrian, D. (2025). Comparative Analysis of SDLC and R&D Methods in System Development: A Case Study of Integrity Zone Management System. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, 9(4), 3197–3209. <https://doi.org/10.33395/SINKRON.V9I4.15337>
- Perdana, A., Farhana, N. A., Harliana, P., Muslim, I., & Karo, K. (2024). Web-Based Application Development using PHP-Native Framework on Agent of Change Integrity Zone Information System. *Jurnal.Polgan.Ac.IdA Perdana, NA Farhana, P Harliana, IMK KaroSinkron: Jurnal Dan Penelitian Teknik Informatika, 2024•jurnal.Polgan.Ac.Id, 8(4)*. <https://doi.org/10.33395/sinkron.v8i4.14118>
- Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi. (2019). *Peraturan Menteri PANRB Nomor 10 Tahun 2019 tentang Perubahan atas Peraturan Menteri PANRB Nomor 52 Tahun 2014 tentang Pedoman Pembangunan Zona Integritas Menuju Wilayah Bebas dari Korupsi dan Wilayah Birokrasi Bersih dan Melayani di Lingkungan Instansi Pemerintah*. <https://peraturan.bpk.go.id/Details/143766/permen-panrb-no-10->
- Riansyahputra, M., Aziz, A., & Purwanto, A. (2025). X Sentiment Analysis on Indonesia's New Capital (IKN) Using TF-IDF+SVM and IndoBERT, and Its Policy-Monitoring Implications. *ICoBITS, 1*, 638–651. <https://doi.org/10.32664/ICOBITS.V1.106>
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, 5(1). <https://doi.org/10.1007/s41133-020-00032-0>
- Singh, D., Bhatnagar, M., & Yadav, V. (2022). PDF Classification Using Logistic Regression and Latent Dirichlet Allocation. *Lecture Notes in Networks and Systems*, 237, 399–407. https://doi.org/10.1007/978-981-16-6407-6_36/SAVE-RESEARCH
- Tarekegn, A. N., Ullah, M., & Cheikh, F. A. (2024). *Deep Learning for Multi-Label Learning: A Comprehensive Survey*. <https://arxiv.org/pdf/2401.16549>
- Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Fung, P., Baldwin, T., Lau, J. H., Sennrich, R., & Ruder, S. (2023). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 815–834. <https://doi.org/10.18653/V1/2023.EACL-MAIN.57>