



CVI-Validated Indo-Transformer Framework for Intelligent Cooperative Supervision

Syahroni Hidayat¹, Afriani Fajar Navissaturrisqi^{2*}, Feddy Setio Pribadi³

¹Department of Electrical Engineering, Universitas Negeri Semarang, Indonesia

²Department of Informatics and Computer Engineering Education, Universitas Negeri Semarang, Indonesia

³Department of Computer Engineering, Universitas Negeri Semarang, Indonesia

*afrianinavis@students.unnes.ac.id

Abstract

Cooperative supervision reports contain complex narrative structures and overlapping administrative terminology, complicating automatic classification into governance, risk profile, financial performance, and capital adequacy. Reliable automation is particularly important for accelerating the analysis of supervisory findings while addressing limited labeled data and imbalanced categories. This study aimed to develop and externally evaluate a text-classification framework combining quantitatively validated Generative Artificial Intelligence (GenAI) labeling with conventional and Transformer-based models. Data comprised 294 preprocessed sentences collected from the Department of Cooperatives, Small and Medium Enterprises, Industry, and Trade of Semarang Regency during 2023–2025. Few-shot annotations were generated using ChatGPT, Gemini, Perplexity, and DeepSeek, and three-model combinations were evaluated using the Content Validity Index (CVI); majority voting from the best combination established ground truth. TF-IDF with Logistic Regression and Support Vector Machine served as baselines, whereas IndoBERT and IndoRoBERTa represented contextual models. Performance was assessed through stratified five-fold cross-validation and external testing on 58 unseen sentences. ChatGPT–Gemini–Perplexity achieved the highest Scale-Level CVI of 0.898. IndoBERT obtained the best cross-validated F1-score of 0.9099, exceeding IndoRoBERTa (0.8217), Logistic Regression (0.8004), and SVM (0.7863). On unseen data, IndoBERT retained an F1-score of 0.862, compared with 0.759 for IndoRoBERTa. These findings demonstrate that CVI-validated ensemble GenAI can construct consistent labels for low-resource administrative texts and that IndoBERT provides the strongest and most stable generalization for cooperative supervision classification. The framework offers a practical basis for scalable annotation and reliable automated support for evidence-based supervisory decision-making.

Keywords: cooperative supervision; GenAI; IndoBERT; stratified k-fold; text classification

How to Cite: S. Hidayat, A. F. Navissaturrisqi, and F. S. Pribadi, “CVI-Validated Indo-Transformer Framework for Intelligent Cooperative Supervision”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 1013 - 1024, Aug. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7606>

Received: February 26, 2026

Accepted: July 16, 2026

Available Online: August 4, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Cooperatives constitute a fundamental pillar of the people-centered economy in Indonesia. They are founded upon principles of kinship and mutual cooperation [1]. To ensure the sustainability of their economic functions and regulatory compliance, the government through the Regulation of the Minister of Cooperatives and SMEs (Permenkop) No. 9 of 2020 has established stringent supervisory standards [2]. This supervisory framework encompasses governance, risk profile, financial performance, and capital adequacy aspects [3]. However, the implementation of these supervisory mechanisms has generated new challenges, particularly the accumulation of unstructured

administrative documents, especially in the form of Official Supervisory Reports (*Berita Acara Hasil Pengawasan*). These documents contain critical findings written in free-text narratives with high variability in terminology and writing styles. Such semantic ambiguity hinders automatic categorization into governance, risk profile, financial performance, and capital adequacy aspects [4]. Ultimately, this ambiguity delays strategic decision-making processes.

In efforts to automate text classification, conventional Machine Learning approaches such as Support Vector Machine (SVM) and Logistic Regression with Term Frequency Inverse Document Frequency (TF-IDF) feature representations are commonly employed as

baseline models. Although these methods offer high computational efficiency, they exhibit fundamental limitations in capturing long-range semantic relationships, handling complex contextual information, and processing formal governmental documents with domain-specific linguistic characteristics [5]. In contrast, the emergence of Transformer-based architectures such as IndoBERT and IndoRoBERTa provides more advanced Natural Language Understanding (NLU) capabilities through bidirectional contextual representations trained on large-scale corpora. However, the application of these models to domain-specific cooperative supervisory reports with limited datasets remains underexplored [6]. Therefore, empirical evaluation is required to determine whether the architectural complexity of Transformer models can consistently deliver superior performance compared to simpler baseline algorithms in low-resource settings [7].

Another persistent challenge in administrative-domain NLP research is the scarcity of labeled data (low-resource data), compounded by the difficulty of engaging human experts in time-consuming annotation processes. The utilization of Generative Artificial Intelligence (GenAI) as an expert annotator has recently emerged as a promising and efficient alternative [8]. However, the reliability of labels generated by a single AI model is often questioned. To address this concern, this study introduces a rigorous validation framework employing an ensemble of multiple GenAI models as label providers. The resulting annotations are subsequently evaluated using the Content Validity Index (CVI). The application of CVI traditionally recognized as a standard instrument validation measure in quantitative research to assess AI-generated label consensus represents a novel approach, ensuring that the constructed dataset achieves high content validity comparable to expert human evaluation [9].

Beyond validity concerns, administrative datasets frequently exhibit class imbalance issues. Such imbalance risks biasing classification models toward majority classes, thereby reducing the system's reliability in detecting minority categories that may carry high levels of urgency [10]. Consequently, selecting an appropriate evaluation strategy becomes crucial to ensuring model robustness under data constraints. Stratified k-fold cross-validation is widely acknowledged as an effective method for evaluating small and imbalanced datasets, as it preserves label proportions across each validation fold, yielding more stable and objective performance estimates compared to standard cross-validation techniques [11].

The primary objective of this study is to develop a classification model capable of categorizing sentences describing cooperative supervision findings into four supervision aspects, namely governance, risk profile, financial performance, and capital adequacy. Specifically, this study aims to: (1) implement a scientifically validated dataset labeling method using a

GenAI ensemble approach assessed through the CVI framework; (2) construct and evaluate a Transformer-based classification pipeline (IndoBERT and IndoRoBERTa) in comparison with baseline algorithms (SVM and Logistic Regression) to validate the effectiveness of NLU architectures in administrative documents; and (3) evaluate the performance, robustness, and generalization ability of the IndoBERT and IndoRoBERTa models in handling imbalanced class distributions on a limited-size dataset using a stratified k-fold cross-validation approach, as well as performing predictions on new data to assess model performance.

In line with these objectives, the novelty of this study lies in the proposed ensemble Generative Artificial Intelligence (GenAI) labeling framework, which is validated using the Content Validity Index (CVI) to support the construction of a high-quality labeled dataset. Unlike previous studies that primarily focus on improving classification performance, this study emphasizes the development and validation of a systematic GenAI-based labeling approach for Indonesian cooperative supervision documents, a domain that remains relatively underexplored in previous research. In addition, this study contributes empirical evidence regarding the performance of Transformer-based models on small and imbalanced administrative text datasets, providing insights into their stability, robustness, and generalization capability in real-world cooperative supervision contexts.

2. Methods

2.1 Research Design

This quantitative experimental study compares baseline and Transformer-based models for classifying cooperative supervisory findings. Generative Artificial Intelligence (GenAI) produced data labels, which were validated using the Content Validity Index (CVI) to ensure reliable training and evaluation data. The intelligent system compares SVM and Logistic Regression with IndoBERT and IndoRoBERTa.

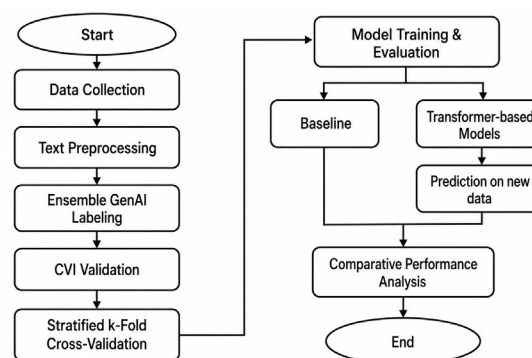


Figure 1. Research Procedure

Models were trained and evaluated through stratified k-fold cross-validation to preserve class proportions and reduce imbalance-related bias. To assess generalization,

the highest-performing model from each Transformer architecture was selected based on cross-validation scores and applied without retraining to unseen data. Figure 1 summarizes the labeling, preprocessing, training, evaluation, and external-prediction workflow.

2.2 Data Collection

The primary data consist of short textual statements describing cooperative weaknesses (5-10 words per sentence), collected from the *Minutes of Cooperative Guidance and Supervision Reports* for the period 2023-2025. The documents were obtained from the Semarang Regency Office of Cooperatives (*Dinas Koperasi Kabupaten Semarang*). Each sentence in the document represents a single finding associated with one of four supervisory aspects: governance, risk profile, financial performance, or capital adequacy. The original documents were handwritten in Bahasa Indonesia. The collected dataset is illustrated in Figure 2.

Examples of cooperative weakness statements include “The cooperative has not conducted an external financial statement audit.”, “The cooperative engages in activities beyond savings and loans, inconsistent with its cooperative type.”, “The management term of office in the articles of association is only two years.”, These short textual findings form the basis of the supervised classification task in this study.



Figure 2. Example of a text dataset on cooperative weaknesses found within the minutes of cooperative guidance and supervision reports

2.3. Preprocessing

Preprocessing standardized the dataset structure for effective model processing and reduced formatting inconsistencies and administrative-language noise that could influence GenAI-based labeling [12]. Hardcopy supervisory reports were digitized using Optical Character Recognition (OCR), and outputs were manually checked against the original handwritten documents. Text was then normalized by converting characters to lowercase, except official abbreviations, and simplifying inconsistent formats. Document abbreviations were manually expanded to clarify their meanings. These procedures preceded labeling, ensuring that GenAI models annotated standardized, semantically consistent text.

2.4 GenAI-based Labeling and CVI Validation

Data labeling assigns annotations or tags to raw data, adding contextual information required for supervised learning. Labeled datasets provide input-output pairs through which models learn relationships between features and target classes for accurate classification or prediction [13].

Based on the workflow in Figure 3, this study proposes a quantitatively validated Multi-Model Ensemble Labeling framework using Generative AI. The framework addresses limitations of human annotators in specialized administrative domains and reduces subjectivity associated with single-source labeling. It employs four cross-company LLMs: GPT-4o mini (ChatGPT), Gemini 1.5 Flash, Perplexity’s Sonar model, and DeepSeek-V3. Their architectural and training-data diversity helps mitigate systematic and self-preference biases within individual models [14].

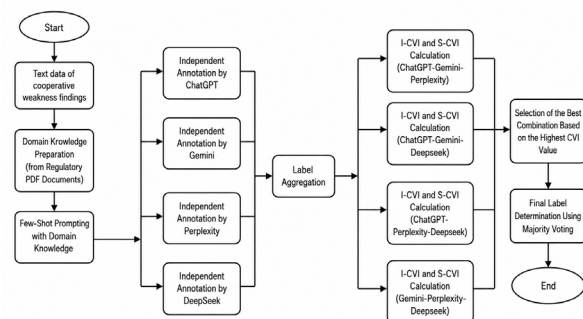


Figure 3. Labeling process flow

Labeling uses few-shot prompting to clarify semantic distinctions among classes. Examples incorporate regulatory knowledge from government regulations and statutory documents defining supervisory aspects. Every model receives identical instructions to standardize annotation outputs. The few-shot labeling instruction used in this study is presented as follows:

“I am conducting a text classification study on findings of weaknesses identified in cooperative supervision reports using Natural Language Processing (NLP) classification models based on machine learning and deep learning approaches. At this stage, the research is focused on data labeling using Generative Artificial Intelligence (GenAI). The dataset consists of sentences extracted from cooperative supervision report documents.

Your task is to classify each sentence into exactly one of the following four categories, Governance (G), Risk Profile (R), Financial Performance (F), and Capitalization (C). Please carefully read and understand the few-shot prompting examples and the knowledge document provided in the appendix above. After understanding the definitions and characteristics of each category, assign the single most appropriate label to each sentence presented below.”

The few-shot prompting template, examples, and knowledge document are provided in Appendices A and B to support transparency and reproducibility. Outputs

from the four GenAI models were arranged into four three-model combinations: ChatGPT–Gemini–Perplexity, ChatGPT–Gemini–DeepSeek, ChatGPT–Perplexity–DeepSeek, and Gemini–Perplexity–DeepSeek. Each combination was independently evaluated using the Content Validity Index (CVI), a measure of rater agreement ranging from 0 to 1; values above 0.80 indicate good, methodologically acceptable content validity [15]. CVI assessment comprised Item-Level CVI (I-CVI) and Scale-Level CVI (S-CVI) [16]. For each item, agreement among the three models was calculated using Equation 1.

$$I - CVI = \frac{n_e}{N} \quad (1)$$

n_e denotes the number assigning an identical label and N is the number of models in the combination. Overall agreement was calculated using Equation 2.

$$S - CVI = \frac{\sum I - CVI}{m} \quad (2)$$

$\sum I - CVI$ denotes the sum of all item-level CVI values, and m is the total number of validated items.

where m denotes the total number of validated items. I-CVI served as the primary item-level consensus criterion. Labels with I-CVI ≥ 0.67 were retained because at least two of three models agreed, whereas items below 0.67 were removed to preserve data integrity. S-CVI assessed overall agreement for each annotator combination. The combination achieving the highest S-CVI was selected for final label assignment, ensuring both item-level validity and global consistency. Within this best-performing combination, the majority label was designated as ground truth only when at least two of the three GenAI models agreed. This procedure reduces dependence on any single model and combines local consensus filtering with selection of the most globally consistent annotator set, thereby producing a reliable dataset for subsequent supervised-learning analysis and downstream model development.

2.5 Text Representation and Classification Model Pipelines

This study employs two text-classification pipelines to categorize cooperative weakness findings into governance, risk profile, financial performance, and capital adequacy. The first pipeline combines TF-IDF features with Logistic Regression and Support Vector Machine (SVM) baselines. The second tokenizes text and fine-tunes IndoBERT and IndoRoBERTa, with Optuna optimizing their hyperparameters. These pipelines compare traditional statistical methods and Transformer architectures on small, class-imbalanced datasets. TF-IDF baselines provide frequency-based representations and are often more robust and computationally efficient than complex deep-learning models under low-resource conditions [17].

Conversely, IndoBERT and IndoRoBERTa provide contextual semantic representations capable of capturing linguistic ambiguity and long-range

dependencies in Indonesian text that TF-IDF may inadequately represent [18]. Optuna reduces overfitting risk by efficiently automating hyperparameter-space exploration and can outperform conventional Grid Search or Random Search [19].

Although Transformers offer stronger semantic representations, their small-dataset performance depends substantially on fine-tuning; consequently, comparison with classical models remains necessary to assess prediction stability under class imbalance [20]. All models were evaluated using stratified k-fold cross-validation, ensuring that every fold retained the original class proportions. Stratification mitigates evaluation bias because conventional cross-validation can generate partitions with insufficient minority-class representation [21], thereby improving the reliability and comparability of performance estimates across all evaluated classifiers.

Logistic Regression and Support Vector Machine (SVM) were used as benchmarks for evaluating Transformer-based text classifiers. With TF-IDF features, these models capture frequency patterns in short, formal Indonesian administrative texts [22], [23]. Their inclusion verifies that performance improvements are meaningful rather than dataset-specific variation, while demonstrating the efficiency and robustness of conventional methods on small or imbalanced datasets [24], [25]. Texts were converted into TF-IDF vectors using unigram and bigram features, preserving individual terms and limited contextual relationships. Term importance was calculated using Equation 3.

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{df(t)}\right) \quad (3)$$

N is the corpus size and $df(t)$ is the number of documents containing term t [26]. The feature space was restricted to the 5,000 highest-ranked features to improve efficiency and reduce dimensionality.

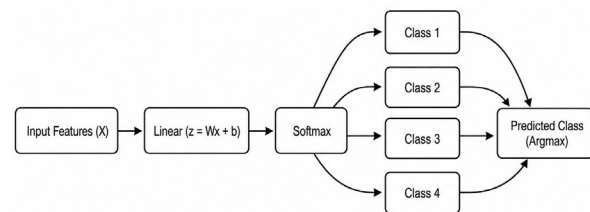


Figure 4. The architecture of the Multinomial Logistic Regression model for multi-class classification.

As illustrated in Figure 4, Logistic Regression linearly combines input features x with weights w and bias b . Although binary Logistic Regression commonly applies a sigmoid function [27].

Figure 5 presents the SVM architecture [28]. SVM constructs a hyperplane that maximizes the distance to the nearest observations, termed support vectors [29]. LinearSVC was implemented through a One-vs-Rest strategy, producing four hyperplanes, each separating one class from the remaining classes. Setting `class_weight='balanced'` assigned greater importance to

minority classes, maintaining effective margins despite class imbalance. Together, these baselines provide interpretable, computationally efficient reference results for assessing whether contextual Transformer representations offer substantive improvements across the four supervisory aspects evaluated.

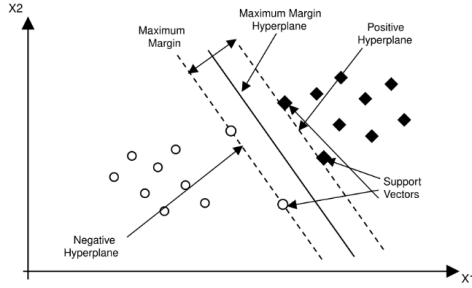


Figure 5. Architecture of support vector machine (SVM)

IndoBERT and IndoRoBERTa were selected because BERT-based models often outperform traditional classifiers [30], while self-attention captures contextual relationships [31]. Preprocessed texts were tokenized using WordPiece for IndoBERT and Byte-Pair Encoding for IndoRoBERTa [27], both decompose rare or out-of-vocabulary terms into subwords [28]. Optuna optimized learning rate, batch size, weight decay, and warmup ratio through 10 Tree-structured Parzen Estimator trials per model [32].

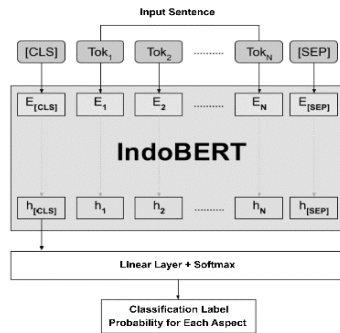


Figure 6. Architecture of IndoBERT

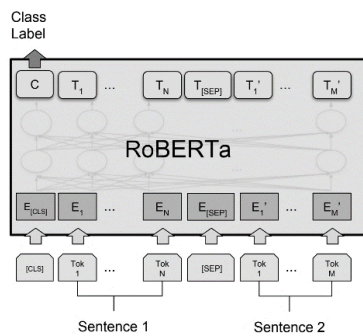


Figure 7. Architecture of RoBERTa

Configurations were assessed using mean F1-score under five-fold stratified cross-validation. IndoBERT (Figure 6) combines multi-head self-attention and feed-forward encoders; Masked Language Modeling produces bidirectional representations [33], [34]. Its fine-tuned base variant classifies four categories using

the [CLS] representation. IndoRoBERTa (Figure 7) removes Next Sentence Prediction, applies dynamic masking [35], [36], and similarly classifies through [CLS]. The 294 instances were divided into five proportionally stratified folds, using approximately 80% for training and 20% for validation. This design preserved minority-class representation and supported comparison of average generalization across both optimized Transformer architectures.

To assess generalization comprehensively, the study evaluates new cooperative supervision and inspection reports excluded from training and validation. These reports undergo the original preprocessing pipeline, including OCR-based digitization, text normalization, and GenAI-based labeling. Ground-truth labels are produced by the three-model GenAI combination achieving the highest S-CVI on the training dataset, with majority voting assigning the most frequent label. The best IndoBERT and IndoRoBERTa models, defined as the folds obtaining the highest F1-scores during stratified k-fold cross-validation, then predict these unseen reports. Both models result from fine-tuning with Optuna-optimized hyperparameters and are applied directly without retraining. New texts are tokenized according to their respective architectures: WordPiece for IndoBERT and Byte-Pair Encoding for IndoRoBERTa. Model predictions are subsequently compared with the GenAI-derived ground truth to quantify agreement on entirely unseen data. Accuracy, precision, recall, and F1-score provide a comprehensive evaluation of classification performance under more realistic unseen-data scenarios and real-world deployment conditions.

2.6 Evaluation Metrics

Each model was evaluated on classifying cooperative weakness findings into four supervisory aspects using accuracy, precision, recall, and F1-score, which collectively support assessment under class imbalance [37]. These metrics were calculated using Equations 4 – 7.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy represents the proportion of correct predictions; precision measures the reliability of positive predictions by penalizing false positives; recall quantifies the completeness of detecting actual positive cases; and F1-score provides their harmonic balance. Metric values from all stratified k-fold cross-validation folds were averaged to produce final scores for comparing baseline machine-learning and Transformer-based approaches. Additionally, the standard deviation

of fold-level F1-scores was calculated to assess stability across data partitions. A low standard deviation indicates consistent performance and limited sensitivity to training-data variation, whereas a high value indicates substantial performance variability among evaluated folds.

3. Results and Discussions

3.1 Label Validation and Data Distribution

The labeling process employed an ensemble Generative Artificial Intelligence (GenAI) framework in which multiple large language models collectively annotated Indonesian administrative texts. Beyond automating annotation, this approach examined semantic consensus across models, while the Content Validity Index (CVI) validated annotation quality and consistency.

Among four three-model combinations, ChatGPT, Gemini, and Perplexity achieved the highest Scale-Level CVI (S-CVI), 0.898, and were therefore selected to determine final labels. Table 1 reports all combination-level S-CVI values, while Appendix C provides complete dataset-level labeling results. Because 0.898 exceeds the minimum acceptable threshold of 0.80 [14], the labeling process demonstrated good content validity.

Table 1. CVI calculation results for each GenAI combination

Combination	S-CVI
ChatGPT, Gemini, Perplexity	0.898
ChatGPT, Gemini, DeepSeek	0.854
ChatGPT, Perplexity, DeepSeek	0.852
Gemini, Perplexity, DeepSeek	0.882

Of 300 sentences, 215 achieved an Item-Level CVI (I-CVI) of 1.00, 79 achieved 0.67, and six achieved 0.33. Thus, 71.6% reached perfect agreement, indicating strong intermodel consistency in category identification. This agreement demonstrates that the ensemble framework captures stable semantic patterns despite differences in model architectures and training corpora. Consensus therefore emerged from aggregating multiple model decisions under a uniform regulation-based prompting strategy rather than depending on one model.

CVI functioned as a quantitative validation layer, not the core labeling mechanism. It identified labels with acceptable agreement and reduced noisy annotations commonly associated with single-model labeling. The six sentences obtaining I-CVI=0.33 reflected complete disagreement, because each model assigned a different category. These sentences failed the predefined consensus threshold of $I-CVI \geq 0.67$ and were excluded. Consequently, the final dataset contained 294 sentences with sufficient majority agreement for subsequent model development and evaluation. This filtering procedure retained all items supported by at least two annotators, preserving usable data while preventing unresolved semantic ambiguity from entering training and ensuring that downstream comparisons were based

on labels supported by transparent, quantitatively established intermodel consensus evidence.

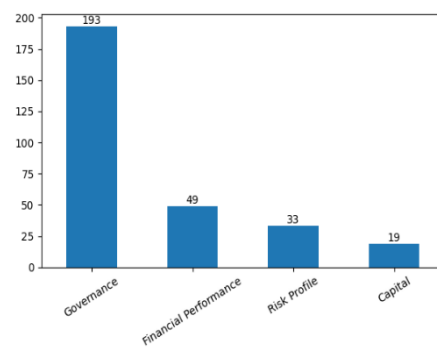


Figure 8. Data distribution across supervision aspects

Table 2. Percentage Distribution of Data Across Aspects

Aspect	Number of Sentences	Percentage
Governance	193	65.65%
Financial Performance	49	16.67%
Risk Profile	33	11.22%
Capitalization	19	6.46%
Total	294	100%

Figure 8 shows that governance dominates the final dataset with 193 sentences, followed by financial performance (49), risk profile (33), and capitalization (19). Table 2 reports their percentages. This distribution reveals substantial imbalance, with governance exceeding half the dataset and capitalization representing the smallest class. Stratified k-fold cross-validation was therefore required to preserve class proportions and improve evaluation representativeness, particularly for minority classes. Although LLMs may generate convincing but inaccurate information [14], the ensemble framework cross-verifies multiple outputs to reduce individual-model bias [38]. Stable S-CVI values across combinations demonstrate that consensus arises from collective decisions rather than one model. This finding supports Seo's use of LLMs as evaluators [39] while extending it through CVI validation of Indonesian administrative texts. The framework supports reliable semi-automated construction under annotator scarcity.

3.2 Baseline Model Evaluation Using Lexical Features

Baseline models combined TF-IDF unigram and bigram features with Logistic Regression or Support Vector Machine (SVM). Stratified five-fold cross-validation was used, and fold-level results were averaged as shows in Table 3. Logistic Regression achieved accuracy, precision, recall, and macro F1-score values of 0.8743, 0.8525, 0.7804, and 0.8004, respectively, slightly outperforming SVM at 0.8641, 0.8512, 0.7562, and 0.7863.

Nevertheless, Figure 9 shows more frequent errors in minority classes, particularly risk profile, which was sometimes predicted as governance. This pattern reflects class imbalance and potential majority-class bias [40], reinforcing the importance of stratified cross-validation for preserving class proportions and

evaluation reliability [41]. It also indicates TF-IDF's dependence on lexical similarity.

Table 3. Average evaluation across five folds for TF-IDF + Logistic Regression and TF-IDF + SVM.

Model	Accuracy	Precision	Recall	F1-Score
TF-IDF + Logistic Regression	0.8743	0.8525	0.7804	0.8004
TF-IDF + SVM	0.8641	0.8512	0.7562	0.7863

Administrative texts use formal, broadly shared terms, including “supervision,” “compliance,” “report,” and “evaluation,” across categories with different contextual meanings. Because unigram–bigram TF-IDF represents frequency and co-occurrence without broader contextual relationships, it struggles with semantic disambiguation when terminology overlaps. Risk profile consequently becomes difficult to distinguish from governance and financial performance. These results demonstrate the limitations of frequency-based models in administrative domains and motivate contextual representations that model inter-token relationships and subtle conceptual differences more comprehensively.

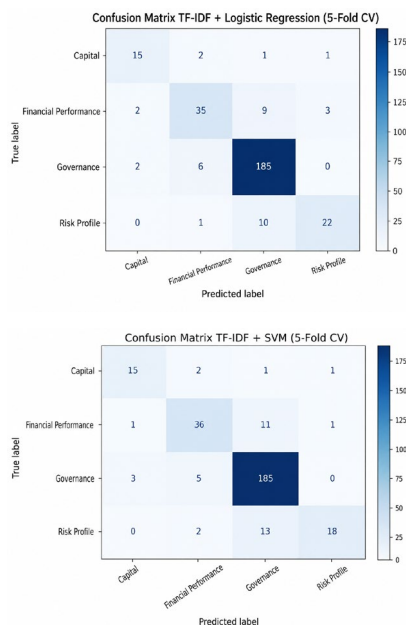


Figure 9. Confusion matrices of a)TF-IDF + Logistic Regression and TF-IDF + SVM

3.3 Transformer Model Evaluation and Analysis

This section evaluates IndoBERT and IndoRoBERTa on the cooperative supervisory dataset using stratified k-fold cross-validation. Optuna optimized the hyperparameters to reduce overfitting on the small, imbalanced dataset. Accuracy, precision, recall, and F1-score were calculated for each fold and averaged for comparison with the baseline models.

Hyperparameters for IndoBERT and IndoRoBERTa were optimized with Optuna before final training, using stratified five-fold cross-validation to preserve class proportions. Ten trials explored learning rates from 1e-5 to 5e-5, 2–3 epochs, batch sizes of 8 or 16, weight

decay from 1e-6 to 1e-3, and warmup ratios from 0.0 to 0.1. For IndoBERT, the optimal configuration reported in Table 4 achieved a macro F1-score of 0.82 based on Figure 10. Figure 10 shows that this value emerged early, while later trials fluctuated considerably and failed to improve the best-value curve. IndoBERT therefore reached a near-optimal search region quickly but remained sensitive to hyperparameter variation.

Initial IndoRoBERTa trials using 2–3 epochs achieved only 0.19 macro F1, suggesting insufficient fine-tuning for its larger capacity, the administrative domain, and class imbalance. Expanding the search to 4–5 epochs produced the optimal configuration in Table 4 and increased macro F1 to 0.77 based on Figure 10. Figure 11 shows gradual, stepwise improvement, reaching approximately 0.77 at trial 6 before stabilizing, although intermediate trials fluctuated. IndoRoBERTa consequently required longer adaptation and broader epoch exploration than IndoBERT. These patterns must be interpreted considering the dataset's 294 labeled sentences and long-tail distribution. Even with stratification, small differences in fold composition can alter class representation, gradient updates, and validation outcomes.

Table 4. Best hyperparameter configuration of IndoBERT and IndoRoBERTa obtained using Optuna.

	Learning Rate	Epoch	Batch Size	Weight Decay	Warmup Ratio
IndoBERT	3.54e-5	3	8	0.0006	0.0737
IndoRoBERTa	3.89e-5	5	16	0.00082	0.0835

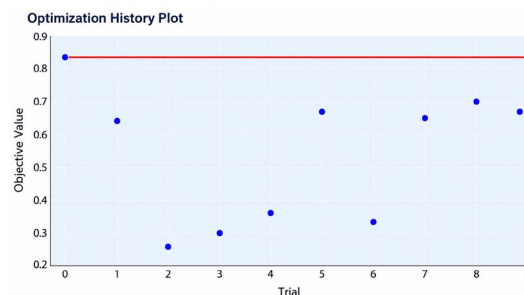


Figure 10. Optimization History Plot of IndoBERT

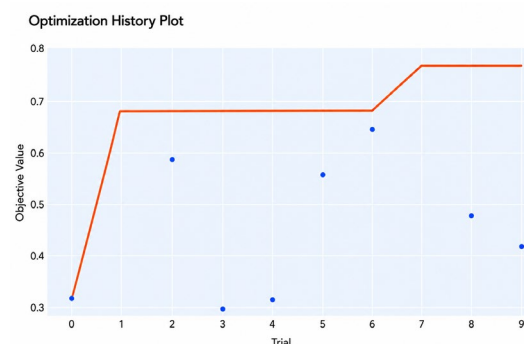


Figure 11. Optimization History Plot of IndoRoBERTa

Minor changes in learning rate, epochs, or other settings may therefore produce disproportionately large performance differences. Accordingly, fluctuations in Figures 10 and 11 reflect interactions between model

capacity and limited data rather than Transformer architecture alone. Their generalizability to larger or more balanced datasets should consequently be treated cautiously. Nevertheless, the selected configurations provide the strongest cross-validated settings identified within the predefined search spaces and were subsequently used for final model training and comparative evaluation in this study.

Following Optuna optimization, the best configurations were applied to IndoBERT and IndoRoBERTa using stratified five-fold cross-validation. Although optimization selected three epochs for IndoBERT and five for IndoRoBERTa, these values came from restricted search spaces and served only as guidance. Final training therefore allowed up to 20 epochs, while early stopping dynamically terminated learning when validation performance ceased improving. This strategy provided sufficient convergence time while limiting premature termination and overfitting on the small dataset [42]. Figure 12 presents average training and validation losses across folds.

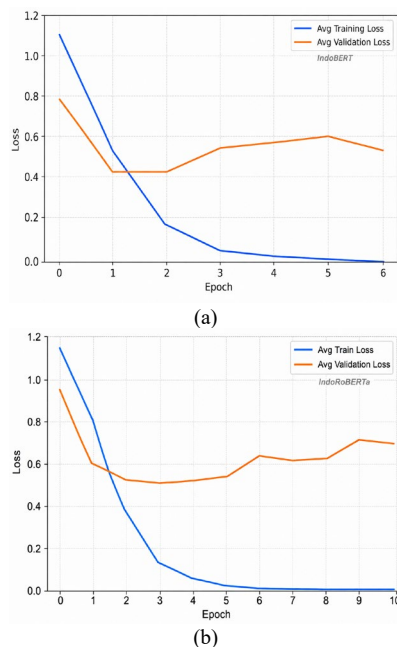


Figure 12. Average training and validation loss of a) IndoBERT and b) IndoRoBERTa

Both models showed consistently declining training loss, approaching zero near epoch 5. For IndoBERT, validation loss initially reached approximately 0.43 but subsequently increased toward 0.60, indicating overfitting after early epochs. IndoRoBERTa's validation loss decreased only slightly before fluctuating between 0.58 and 0.62, despite its training loss approaching zero. Thus, greater model capacity did not improve generalization under limited data.

The widening training–validation loss gap reflects the bias–variance tradeoff [43], complex models trained on small datasets exhibit greater variance, and additional epochs may intensify overfitting. This interpretation is consistent with final metrics and confusion matrices showing that larger capacity does not necessarily

improve minority-class robustness. Accordingly, fine-tuning administrative text models under long-tail, low-resource conditions requires careful control of model complexity to balance learning capacity and generalization.

Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices under stratified five-fold cross-validation, preserving class proportions across folds [41]. Table 5 shows that IndoBERT achieved a macro F1-score of 0.9099, outperforming IndoRoBERTa at 0.8217 and demonstrating greater stability under long-tail conditions [40].

Table 5. Average evaluation results across 5-fold cross-validation for transformer models

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT	0.9116	0.9171	0.9116	0.9099
IndoRoBERTa	0.8877	0.8589	0.7956	0.8217

Figure 13 shows strong governance classification, with 189 correct predictions for IndoBERT and 185 for IndoRoBERTa. Corresponding correct predictions were 36 versus 37 for financial performance, 17 versus 16 for capital, and 26 versus 20 for risk profile. Errors concentrated in minority classes, particularly risk profile, which was sometimes classified as governance because of overlapping technical terminology. Self-attention enables contextual disambiguation beyond frequency-based representations [44], although greater complexity does not guarantee robustness on limited data, consistent with the bias–variance tradeoff [43].

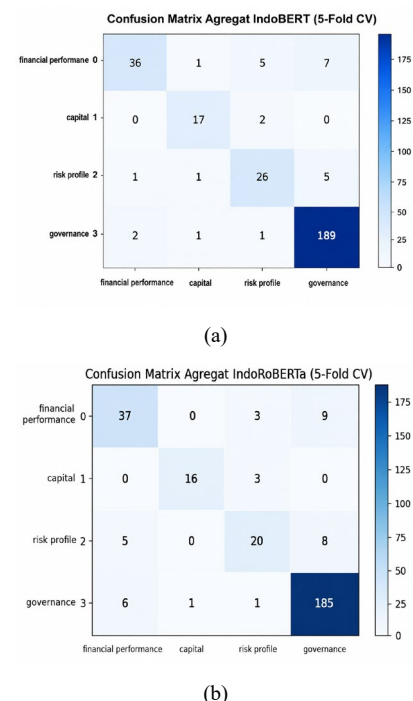


Figure 13. confusion matrix of a) IndoBERT and b) IndoRoBERTa

The study remains limited by its small, imbalanced dataset from one local government institution. Moreover, using the same folds for Optuna optimization and evaluation may introduce optimistic

bias. Future work should collect multi-institutional data, apply nested cross-validation to separate optimization from evaluation, and investigate additional imbalance-handling strategies to improve robustness, generalizability, and practical applicability. These measures would provide more realistic estimates of external model performance.

3.3.4 Prediction Results of New Data Labels Using IndoBERT and IndoRoBERTa Models

External generalization was assessed using 61 sentences from cooperative development and supervision reports excluded from training and validation. Each sentence underwent the original preprocessing pipeline, including OCR digitization, text normalization, and ensemble GenAI labeling. ChatGPT, Gemini, and Perplexity, the combination previously achieving the highest S-CVI, assigned labels through majority voting. Three sentences received three different predictions and were excluded; therefore, 58 sentences with agreement from at least two models formed the external test set. Figure 14 shows an imbalanced distribution: governance comprised 26 instances, financial performance 17, risk profile 10, and capitalization 5.

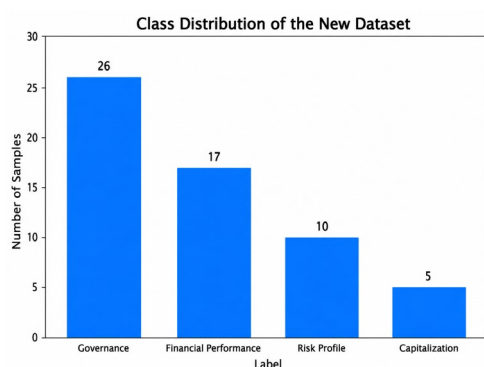


Figure 14. Label Distribution of the New Data

For IndoBERT, Fold 1 weights were selected because this fold achieved the highest internal F1-score, 0.934. On external data, IndoBERT obtained accuracy, precision, recall, and F1-score values of 0.862, 0.873, 0.862, and 0.862, respectively shows in Table 6. Figure 15 shows that most predictions were correctly positioned on the confusion-matrix diagonal. Remaining errors primarily involved financial performance and risk profile. For example, “the cooperative does not have protection against the risk of losing business assets” was labeled as risk profile but predicted as financial performance, probably because “business assets” carries financial associations.

Table 6. Evaluation Results of Predictions on New Data Using the IndoBERT and IndoRoBERTa Model

Evaluation Metric	IndoBERT	IndoRoBERTa
Accuracy	0.862	0.775
Precision	0.873	0.803
Recall	0.862	0.747
F1-Score	0.862	0.759

For IndoRoBERTa, Fold 5 weights, which achieved an internal F1-score of 0.863, were used. External accuracy, precision, recall, and F1-score were 0.775, 0.803, 0.747, and 0.759, respectively shows in Table 6. Thus, precision exceeded recall, indicating that several relevant instances remained unidentified. Figure 16 shows excellent governance recognition, with 25 of 26 instances correct, but weaker minority-class performance: only 9 of 17 financial-performance and 7 of 10 risk-profile instances were correctly classified. Most errors shifted toward governance, the majority class.

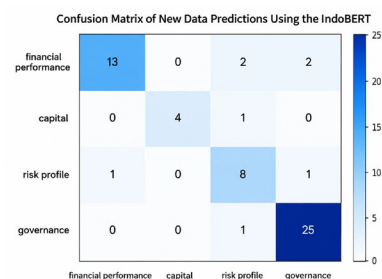


Figure 15. Confusion Matrix of New Data Predictions Using the IndoBERT Model

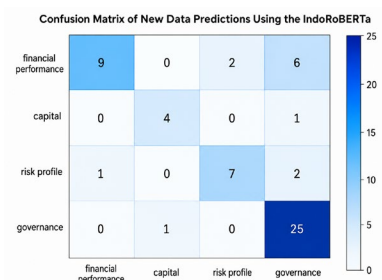


Figure 16. Confusion Matrix of New Data Predictions Using the IndoRoBERTa Model

Although RoBERTa uses dynamic masking and intensive pre-training to develop richer representations, its larger capacity requires sufficient data. The limited, imbalanced training set constrained robust minority-class learning, causing predictions for unfamiliar sentences to favor governance. Consequently, IndoRoBERTa’s F1-score declined from 0.863 internally to 0.759 externally, revealing weaker generalization than suggested by cross-validation. Overall, IndoBERT generalized more effectively, although both models remained vulnerable to semantic overlap and distributional imbalance in real-world administrative data.

These findings demonstrate the importance of external testing because performance estimated from internal folds may not fully represent deployment conditions. IndoBERT retained balanced performance across the unseen dataset, whereas IndoRoBERTa displayed stronger majority-class bias and lower sensitivity to minority categories. The error patterns show that lexical cues can obscure supervisory distinctions when phrases occur across categories. Accordingly, expanding the training corpus, particularly for financial performance, risk profile, and capitalization, is necessary to improve

robustness. Multi-institutional reports and imbalance-mitigation strategies could further strengthen generalization to cooperative supervision documents.

This subsection compares GenAI ground-truth annotations with IndoBERT and IndoRoBERTa predictions to assess agreement and identify interpretative differences. Cooperative-supervision weaknesses were classified into Governance (G), Risk Profile (R), Financial Performance (F), and Capitalization (C). Representative results appear in Table 7, with complete outputs in Appendix D. Agreement was high, particularly for structurally and administratively oriented Governance and Capitalization cases. Discrepancies mainly involved ambiguous sentences. For “the cooperative lacks protection against the risk of business asset loss,” GenAI assigned Risk Profile, whereas both models predicted Financial Performance, indicating an association between asset protection and financial rather than operational risk.

Table 7. Labeling Results between GenAI, IndoBERT, and IndoRoBERTa on New Data

Dataset	Gen AI	Indo BERT	IndoRo BERTa
The cooperative’s organizational structure has not yet been displayed at its office	G	G	G
Members’ principal savings are insufficient to meet the cooperative’s business capital requirements	C	C	C
The cooperative lacks protection against the risk of business asset loss	R	F	F
The cooperative’s business independence remains dependent on government programs	F	R	G

*Governance (G), Risk Profile (R), Financial Performance (F), and Capitalization (C)

For “The cooperative’s business independence remains dependent on government programs,” GenAI assigned Financial Performance, IndoBERT predicted Risk Profile, and IndoRoBERTa predicted Governance, reflecting different interpretations of institutional independence and governmental dependency. Overall, IndoRoBERTa labels aligned more closely with GenAI than IndoBERT labels, especially in structural and administrative categories, suggesting stronger contextual representation of Indonesian cooperative-domain text.

3.4 Comparative Analysis Across Models

Table 8 summarizes five-fold cross-validation performance. TF-IDF with Logistic Regression achieved accuracy, precision, recall, F1-score, and standard deviation values of 0.8743, 0.8525, 0.7804, 0.8004, and 0.0902, respectively; corresponding SVM values were 0.8641, 0.8512, 0.7562, 0.7863, and 0.0913. IndoBERT performed best, reaching 0.9116 accuracy, 0.9171 precision, 0.9116 recall, 0.9099 F1-score, and 0.0137 standard deviation. IndoRoBERTa obtained 0.8877, 0.8589, 0.7956, 0.8217, and 0.0636, respectively.

IndoBERT exceeded Logistic Regression and SVM in F1-score by 0.1095 and 0.1236, confirming the effectiveness of contextual representations for formal administrative texts with overlapping terminology [45]. Unlike TF-IDF, which cannot adequately model semantic token relationships, Transformer self-attention improved cross-class balance. IndoBERT’s lowest standard deviation indicated highly consistent fold-level performance. IndoRoBERTa was more partition-sensitive, while baseline deviations above 0.09 showed greater susceptibility to distributional changes.

Table 8. Average evaluation results across 5-fold cross-validation for each model and method and F1-Score Dttandard Deviation (SD).

Model	Accuracy	Precision	Recall	F1-Score	SD
TF-IDF + Logistic Regression	0.87	0.85	0.78	0.80	0.09
TF-IDF + SVM	0.86	0.85	0.75	0.78	0.09
IndoBERT	0.91	0.91	0.91	0.90	0.01
IndoRo-BERTa	0.88	0.85	0.79	0.82	0.06

Confusion matrices further revealed majority-class bias. Logistic Regression misclassified nine financial-performance and ten risk-profile instances as governance; SVM misclassified eleven and thirteen, respectively. IndoBERT correctly classified 26 risk-profile instances, compared with 20 for IndoRoBERTa, and distributed its errors more evenly. Nevertheless, Transformers still confused some financial-performance cases with governance because shared administrative terms, including “report,” “evaluation,” and “supervision,” carry different contextual meanings.

Overall, Transformer fine-tuning with stratified validation delivered more robust performance on the limited, imbalanced dataset. These results show that higher architectural complexity alone does not guarantee stability: despite its strong accuracy, IndoRoBERTa fluctuated more than IndoBERT. The combined average and variability analysis therefore support IndoBERT as the most reliable model for balanced cooperative-supervision classification across changing folds within this small imbalanced dataset.

4. Conclusions

This study demonstrates that an ensemble Generative Artificial Intelligence approach combining ChatGPT, Gemini, and Perplexity with few-shot prompting achieves robust inter-model agreement. Validated via the Content Validity Index, this framework yielded a high S-CVI score of 0.898, confirming that the ground truth labels are objective and completely independent of a single model. In the classification phase, advanced Transformer-based models, specifically IndoBERT and IndoRoBERTa, significantly outperformed traditional machine learning algorithms such as Logistic Regression and Support Vector Machines across all evaluation metrics, including accuracy, precision,

recall, and F1-score. This clear superiority underscores the immense value of contextual text representations when interpreting complex Indonesian administrative documents filled with overlapping terminology, even when trained on relatively small baseline datasets.

Furthermore, these Transformer architectures provided substantial performance gains for underrepresented minority classes. By utilizing stratified k-fold cross-validation, the experiment successfully maintained consistent class proportions, effectively mitigated majority-class bias, and enhanced empirical reliability under long-tail distribution conditions. Tests on entirely unseen data proved that both IndoBERT and IndoRoBERTa maintain stable predictive capabilities, showing strong generalization despite minor score reductions.

However, limitations persist, particularly the small sample size gathered from just one regional government office and ongoing imbalances within capital adequacy and risk profile categories. Future research should expand data collection across multiple diverse institutions, deploy robust techniques to manage imbalanced sets, and explore multi-label frameworks. Overall, this study delivers crucial theoretical and practical insights for the digital transformation of cooperative supervision in Indonesia, proving the efficacy of deep learning while addressing critical structural data challenges.

Acknowledgements

The authors sincerely thank the Cooperative, Micro, and Industrial Services Office of Semarang Regency for facilitating the data collection for this research.

References

- [1] No. 25 Undang-Undang, "Undang-Undang RI No. 25 Tahun 1992 Tentang Perkoperasian Indonesia [Law of the Republic of Indonesia No. 25 of 1992 on Cooperatives]," *Peraturan Bpk*, no. 25, pp. 1–57, 1992.
- [2] S. A. Abdul Hakim Muhiddin, "Momentum Hari Koperasi, untuk perkuat partisipasi dan aset koperasi [Cooperative Day Momentum to Strengthen Participation and Cooperative Assets]," *Antara*. Accessed: Nov. 20, 2025. [Online]. Available: <https://www.antaranews.com/berita/4960893/momentum-hari-koperasi-untuk-perkuat-partisipasi-dan-aset-koperasi>
- [3] K. K. dan UKKM, "Peraturan Menteri Koperasi dan UKM Nomor 9 Tahun 2020 tentang Pengawasan Koperasi [Regulation of the Minister of Cooperatives and SMEs No. 9 of 2020 on Cooperative Supervision]," *Peraturan Menteri Koperasi dan UKM Nomor 9 Tahun 2020 tentang Pengawasan Koperasi*, no. 1202, pp. 1–53, 2020.
- [4] I. Chalkidis et al., "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 4310–4330, 2021, doi: <https://doi.org/10.18653/v1/2022.acl-long.297>.
- [5] H. S. Mawar and M. Rahardi, "Sentiment Analysis of the TPKS Law on Twitter: A Comparative Study of Classification Algorithm Performance," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3087–3096, 2025, doi: <https://doi.org/10.30871/jaic.v9i6.11503>.
- [6] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, 2025, doi: <https://doi.org/10.1007/s10462-025-11162-5>.
- [7] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, "NBIAS: A natural language processing framework for BIAS identification in text," *Expert Syst. Appl.*, vol. 237, no. PB, p. 121542, 2024, doi: <https://doi.org/10.1016/j.eswa.2023.121542>.
- [8] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose Opinions Do Language Models Reflect?," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, pp. 29971–30004. <https://dl.acm.org/doi/10.5555/3618408.3619652>
- [9] T. Y. C. Tam et al., "A framework for human evaluation of large language models in healthcare derived from literature review," *npj Digital Medicine*, vol. 7, no. 1, 2024, doi: <https://doi.org/10.1038/s41746-024-01258-7>.
- [10] T. Miftahshudur, H. M. Sahin, B. Grieve, and H. Yin, "A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications," *Remote Sensing*, vol. 17, no. 3, pp. 1–31, 2025, doi: <https://doi.org/10.3390/rs17030454>.
- [11] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, 2023, doi: <https://doi.org/10.3390/s23042333>.
- [12] I. R. Huriyah, A. I. S. Widianingrum, Azlina, and Taslim, "Optimasi Augmentasi Data Berbasis Synonym Replacement pada Klasifikasi Teks Berita Menggunakan Neural Network [Optimization of Synonym-Based Data Augmentation in News Text Classification Using Neural Networks]," *Komputa : Jurnal Ilmiah Komputer dan Informatika*, vol. 14, no. 1, pp. 100–107, 2025, doi: <https://doi.org/10.34010/komputa.v14i1.15339>
- [13] S. Fuchs, J. Schnellbach, H. Wittges, and H. Krcmar, "Human vs. Automated data annotation: Labeling the data set for an ML-driven support ticket classifier," *Data Knowl. Eng.*, vol. 162, p. 102534, 2026, doi: <https://doi.org/10.1016/j.datak.2025.102534>.
- [14] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 30, pp. 1–3, 2023, doi: <https://doi.org/10.1073/pnas.2305016120>.
- [15] D. F. Polit and C. T. Beck, "The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations," *Res. Nurs. Health*, vol. 29, no. 5, pp. 489–497, 2006, doi: <https://doi.org/10.1002/nur.20147>.
- [16] M. S. B. Yusoff, "ABC of Content Validation and Content Validity Index Calculation," *Education in Medicine Journal*, vol. 11, no. 2, pp. 49–54, 2019, doi: <https://doi.org/10.21315/eimj2019.11.2.6>.
- [17] S. P. Octarini, A. Y. Zakiyyah, and K. Purwandari, "Comparison of IndoBERT and SVM Algorithm to Perform Aspect Based Sentiment Analysis using Hierarchical Dirichlet Process," *Engineering, Mathematics and Computer Science Journal (EMACS)*, vol. 7, no. 3, pp. 363–370, 2025, doi: <https://doi.org/10.21512/emacsjournal.v7i3.13493>.
- [18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, 2020, doi: <https://doi.org/10.18653/v1/2020.coling-main.66>.
- [19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631. doi: <https://doi.org/10.1145/3292500.3330701>.
- [20] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the Stability of Fine-Tuning BERT: Misconceptions, Explanations, and Strong Baselines," in *International Conference on Learning Representations (ICLR)*, Vienna (Virtual), 2021. <https://doi.org/10.48550/arXiv.2006.04884>
- [21] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, 2020, doi: <https://doi.org/10.1145/3343440>.
- [22] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset," in *COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent*

- Systems, 2021, pp. 98–107. <https://doi.org/10.48550/arXiv.2308.04037>.
- [23] S. Samidi and D. Fatmawati, "Sentiment Classification of IT Service Feedback via TF-IDF," *CogITO Smart Journal*, vol. 10, no. 2, pp. 403–417, 2024, doi: <https://doi.org/10.31154/cogito.v10i2.701.403-417>.
- [24] V. B. Lestari and C. A. Hutagalung, "Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis of Indonesian-Language Marketplaces Using SVM, Logistic Regression, and Naive Bayes," *Journal of Computer Science and Applications*, vol. 8, no. 1, pp. 22–2025, 2025, doi: <https://doi.org/10.21009/j-koma.v8i1.05>.
- [25] V. Ayumi, M. Purba, and S. Mailana, "Klasifikasi Teks Umpan Balik Kompetensi Sosial Dosen di Perguruan Tinggi Menggunakan Word2Vec dan CNN-1D [Text Classification of Lecturer Social Competency Feedback Using Word2Vec and CNN-1D]," *JSAI (Journal Scientific and Applied Informatics)*, vol. 8, no. 2, pp. 564–569, 2025, doi: <https://doi.org/10.36085/jasai.v8i2.8764>.
- [26] W. Trisnawati and A. Wibowo, "Sentiment Analysis of ICT Service User Using Naive Bayes Classifier and Svm Methods with TF-IDF Text Weighting," *J. Tek. Inform. (JUTIF)*, vol. 5, no. 3, pp. 709–719, 2024, doi: <https://doi.org/10.52436/1.jutif.2024.5.3.1784>.
- [27] S. Atikah Permata Sari, "Deteksi Fraud Kartu Kredit Dengan Logistic Regression, Random Forest, Dan Gradient Boosting [Credit Card Fraud Detection Using Logistic Regression, Random Forest, and Gradient Boosting]," *Jurnal Informatika Terpadu*, vol. 6, no. 1, pp. 29–37, 2025, doi: <https://doi.org/10.54914/jit.v1i1.1826>.
- [28] A. Yenikar, C. N. Babu, and D. J. Hemanth, "Semantic relational machine learning model for sentiment analysis using cascade feature selection and heterogeneous classifier ensemble," *PeerJ Comput. Sci.*, vol. 8, pp. 1–34, 2022, doi: <https://doi.org/10.7717/peerj-cs.1100>.
- [29] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information (Switzerland)*, vol. 15, no. 4, 2024, doi: <https://doi.org/10.3390/info15040235>.
- [30] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352–356, 2023, doi: <https://doi.org/10.47852/bonviewJCCE3202838>.
- [31] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, vol. 11, no. 1, 2024, doi: <https://doi.org/10.1186/s40537-023-00842-0>.
- [32] P. Sridevi, Z. Arefin, and S. I. Ahamed, "An integrated machine learning and hyperparameter optimization framework for noninvasive creatinine estimation using photoplethysmography signals," *Healthcare Analytics*, vol. 7, no. April, p. 100395, 2025, doi: <https://doi.org/10.1016/j.health.2025.100395>.
- [33] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: <https://doi.org/10.11591/eei.v13i5.8032>.
- [34] D. Suhartono, M. R. N. Majiid, and R. Fredyan, *Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia*, vol. 29, no. 16. Springer US, 2024. doi: <https://doi.org/10.1007/s10639-024-12717-9>.
- [35] G. Sri Harsh, Y. Sai Sri Vivek, P. Maneesha, S. K. Rout, S. Ranjith Reddy, and B. K. Sethi, "Automated Interview Evaluation System Using RoBERTa Technology," *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing, IC-CGU 2024*, pp. 1–8, 2024, doi: <https://doi.org/10.1109/IC-CGU58078.2024.10530789>.
- [36] F. Dwi Sukma, R. Wijaya, and A. Romadhony, "Speech to Text Correction for Indonesian Early Marriage Counseling Chatbots Using IndoRoBERTa and Mistral-7B," *Journal on Computing*, vol. 10, no. 1, pp. 72–87, 2025, doi: <https://doi.org/10.21108/indoic.v10i1.9708>.
- [37] Q. Li et al., "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, 2022, doi: <https://doi.org/10.1145/3495162>.
- [38] I. D. Mienye and T. G. Swart, "Ensemble Large Language Models: A Survey," *Information (Switzerland)*, vol. 16, no. 8, 2025, doi: <https://doi.org/10.3390/info16080688>.
- [39] H. Seo et al., "Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy," *Applied Sciences (Switzerland)*, vol. 15, no. 2, pp. 1–35, 2025, doi: <https://doi.org/10.3390/app15020671>.
- [40] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: <https://doi.org/10.1016/j.neunet.2018.07.011>.
- [41] T. Abedin, H. Xu, and S. Uddin, "The impact of K selection in K-fold cross-validation on bias and variance in supervised learning models," *Sci. Rep.*, vol. 16, no. 1, pp. 1–13, 2026, doi: <https://doi.org/10.1038/s41598-026-37247-x>.
- [42] B. M. Hussein and S. M. Shareef, "An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning," *ITM Web of Conferences*, vol. 64, p. 01003, 2024, doi: <https://doi.org/10.1051/itmconf/20246401003>.
- [43] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 32, pp. 15849–15854, 2019, doi: <https://doi.org/10.1073/pnas.1903070116>.
- [44] I. Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. <https://doi.org/10.48550/arXiv.1706.03762>.
- [45] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 1, pp. 1–12, 2022, doi: <https://doi.org/10.1186/s12874-022-01665-y>.