VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i4.8221

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

ABSTRACTIVE SUMMARIZATION FOR INDONESIAN HUMAN TRAFFICKING COURT DECISIONS USING TRANSFORMER MODELS

Faradhita Eka Septiana[1]; Febri Bagus Triwibowo[1]; Galih Wasis Wicaksono[1*]

Informatics Study Program[1]
Universitas Muhammadiyah Malang, Malang, Indonesia[1]
https://umm.ac.id[1]
faradhita06@webmail.umm.ac.id, yourbae21333@webmail.umm.ac.id, galih.w.w@umm.ac.id[*]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Indonesian Supreme Court decisions on Human Trafficking (TPPO) are lengthy and structurally complex, rendering manual review inefficient for legal practitioners. Existing abstractive summarization research for Indonesian text concentrates on news and social media, while no publicly benchmarked XSum-style dataset exists for the Indonesian legal domain. This study has two explicit objectives: (i) an XSum-structured legal summarization dataset is constructed from 404 TPPO decisions, and (ii) four fine-tuned Transformer models (T5 Base Indonesia, mT5 Small, DistilBART CNN, BART Large XSum) are benchmarked against extractive and classical abstractive baselines. The method couples an n8n-based PDF extraction pipeline with CSV-sourced verdict statements as reference summaries, followed by fine-tuning and evaluation using ROUGE-1/2/L and BERTScore F1, complemented by paired bootstrap significance testing (n=10,000). Results show T5 Base Indonesia attains the highest ROUGE-L of 39.49 and BERTScore F1 of 74.82, while mT5 Small achieves the highest ROUGE-1 of 44.97, all significantly outperforming Seq2Seq+Attention (ROUGE-1 27.31) and First-2-Sentences (ROUGE-1 10.86) with p<0.001. The contributions are: an XSum-formatted TPPO dataset, an automated extraction pipeline, and a comprehensive benchmark spanning extractive, classical abstractive, and Transformer-based methods. These findings offer practical benefits for legal document analysis and judicial information retrieval in Indonesia.*

Keywords: *Abstractive Summarization, Court Decisions, Indonesian Legal Documents, Transformer Models, XSum Dataset.*

Intisari—*Putusan Mahkamah Agung Indonesia tentang Tindak Pidana Perdagangan Orang (TPPO) berisi teks panjang dengan struktur kompleks sehingga telaah manual tidak efisien bagi praktisi hukum. Riset peringkasan abstraktif untuk teks berbahasa Indonesia masih didominasi domain berita dan media sosial, sedangkan belum tersedia dataset bergaya XSum untuk domain hukum. Studi ini memiliki dua tujuan eksplisit: (i) dataset peringkasan bergaya XSum dibangun dari 404 putusan TPPO, dan (ii) empat model Transformer (T5 Base Indonesia, mT5 Small, DistilBART CNN, BART Large XSum) di-benchmark terhadap baseline ekstraktif dan abstraktif klasik. Metode menggunakan pipeline ekstraksi PDF berbasis n8n dengan amar putusan dari CSV sebagai ringkasan referensi, diikuti fine-tuning dan evaluasi ROUGE-1/2/L serta BERTScore F1, ditambah uji signifikansi paired bootstrap (n=10.000). Hasil menunjukkan T5 Base Indonesia mencapai ROUGE-L tertinggi 39,49 dan BERTScore F1 74,82, sementara mT5 Small mencapai ROUGE-1 tertinggi 44,97, seluruhnya signifikan mengungguli Seq2Seq+Attention (ROUGE-1 27,31) dan First-2-Sentences (ROUGE-1 10,86) dengan p<0,001. Kontribusi meliputi dataset XSum TPPO, pipeline ekstraksi otomatis, serta benchmark komprehensif yang mencakup metode ekstraktif, abstraktif klasik, dan berbasis Transformer. Temuan ini memberikan manfaat praktis bagi analisis dokumen hukum dan retrieval informasi peradilan di Indonesia.*

Kata Kunci: *Abstractive Summarization, Dokumen Hukum Indonesia, Keputusan Pengadilan, Model Transformer, Dataset XSum.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.8221

INTRODUCTION

Indonesian Supreme Court decisions on Human Trafficking (TPPO) are typified by long, discursive text that interleaves factual descriptions, legal considerations, and verdicts in highly formal technical register. Prior surveys confirm that information density in such long documents obstructs human identification of core content and motivates automatic text summarization [1], [2]. An automated approach that preserves legal meaning while producing concise summaries is therefore essential for practitioners, researchers, and judicial information systems.

Automatic text summarization, a core application of Natural Language Processing (NLP), produces compact document representations while retaining essential information [3], and is broadly divided into extractive and abstractive approaches. Extractive methods rank and copy sentences directly from the source and are valued for stability; abstractive methods paraphrase content and may introduce vocabulary not present in the source, at the cost of a non-trivial hallucination risk [4], [5], [6]. For legal documents, factual precision is paramount because textual errors can misrepresent a verdict's substance and erode downstream trust [7], [8].

General-domain abstractive summarization has matured around CNN/DailyMail and XSum [9], [10], [11], and recent domain-specific efforts include CaseSumm for U.S. case law [12], MILDSum for Indian legal judgments [13] and EurLeXSummarization for European multilingual legal texts [14]. Three gaps for the Indonesian TPPO setting remain unresolved, however, and motivate this study. First, the linguistic–structural profile of Indonesian court decisions differs substantially from these corpora [13]: verdict statements (*amar putusan*) employ a fixed performative register *("MENGADILI", "Menyatakan Terdakwa ... terbukti secara sah dan meyakinkan"*) while legal considerations (*pertimbangan hukum*) use extended discursive reasoning, producing an uncommon text-pair profile in which a terse structured target is grounded in a long narrative source. Second, TPPO specifically was selected because trafficking cases exhibit high structural recurrence of named entities (defendant, victim, locus), monetary amounts, and article references from Law No. 21 of 2007, giving strong signal for benchmarking factuality while remaining nationally significant as a priority crime category. Third, there is no publicly available XSum-style Indonesian legal dataset; prior Indonesian summarization work has focused on news and social media rather than judicial text.

Based on these gaps, this study pursues three explicit research objectives: (RO1) a reproducible XSum-formatted summarization dataset comprising 404 Indonesian TPPO Supreme Court decisions is constructed [15]; (RO2) four fine-tuned Transformer architectures T5 Base Indonesia, mT5 Small, DistilBART CNN, and BART Large Xsum are benchmarked against four extractive and four classical abstractive baselines, using ROUGE-1/2/L, BERTScore F1, and a paired bootstrap significance test; and (RO3) a qualitative and quantitative error analysis is conducted to characterise the specific failure modes most relevant to legal deployment. The corresponding contributions are: (C1) an XSum-structured Indonesian TPPO corpus; (C2) an automated text extraction pipeline for Indonesian judicial PDFs; and (C3) a comprehensive benchmark spanning extractive, classical abstractive, and Transformer-based methods, accompanied by statistical significance testing and quantitative error taxonomy to support reproducible evaluation in Indonesian legal NLP.

MATERIALS AND METHODS

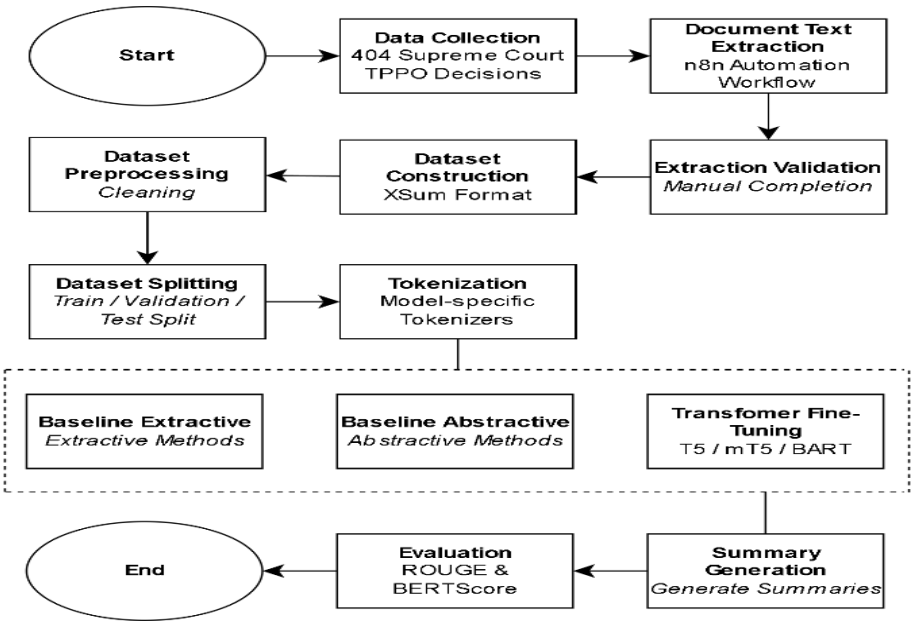
This research develops and evaluates an automatic abstractive summarization system for Indonesian Supreme Court TPPO decisions. The methodology follows a supervised learning paradigm encompassing data collection, PDF extraction, dataset construction, preprocessing, train/validation/test splitting, tokenization, model fine-tuning, and performance evaluation a standard pipeline in supervised abstractive summarization research that enables precise control over summary quality [16], [17], [18].

Research Workflow

The workflow begins with collecting 404 Supreme Court TPPO decisions in PDF format, followed by text extraction using n8n automation and manual validation. The dataset is constructed in XSum format, preprocessed, and split into train/validation/test sets. Experiments compare baseline extractive and abstractive methods against Transformer fine-tuning approaches (T5, mT5, BART), with evaluation using ROUGE, BERTScore, and paired bootstrap tests as summarised in Figure 1.

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i4.8221

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Results, 2026)

Figure 1. Research Workflow of Indonesian Legal Document Summarization

This integrated workflow ensures experimental reproducibility and consistency critical in Transformer-based NLP research where data splitting and experimental settings materially influence conclusions [19], [20], [21].

Data Collection

The dataset comprises 404 Indonesian Human Trafficking court decisions from the Supreme Court's official directory (2023–2024), published on Mendeley Data[15]. Data consists of PDF documents containing complete decision texts and CSV files storing structured verdict statements. This dual-format approach enables clear document-summary separation, following standard practice for explicit document-summary pairing [22].

The sample size of 404 documents was fixed by the population of Supreme Court TPPO decisions released via the official directory during 2023–2024 at the time of data freeze, and is therefore a near-complete snapshot rather than a convenience sample. While small relative to general-domain corpora such as CNN/DailyMail (>300,000), it is comparable to or larger than established legal summarization benchmarks in neighbouring jurisdictions: for example, several language partitions in MILDSum (Indian legal case judgments) [13]and EurLeXSummarization (European multilingual legal texts) [14] operate on sub-thousand instances per language–domain pair. Fine-tuning 220M–406M-parameter encoder–decoders on corpora of this order is also a standard practice for specialised legal and biomedical settings, with regularisation controlled via learning-rate tuning, weight decay, and early stopping (detailed in the Fine-Tuning subsection). Nevertheless, the scope of the corpus remains a constraint on the generalisability of findings, as discussed in the limitations section.

PDF documents contain the formal structure of court decisions including case identity, factual descriptions, legal considerations, and verdicts, along with headers, footers, and watermarks. CSV data directly extracts verdict sections without additional elements. This separation strategy aligns with domain-specific summarization dataset construction practices, where long source documents and reference summaries are explicitly paired for training and evaluating abstractive models, as applied in benchmarks such as CNN/DailyMail, XSum, and XL-Sum [23].

PDF Extraction Process

Text extraction from PDF documents is performed using an n8n-based automation workflow. This approach addresses Indonesian legal documents' inconsistent formats, scanned results, and complex page structures. The n8n system integrates PDF reader modules, initial text cleaning, and structured storage of extraction results, aligning with the information-extraction literature on combining PDF parsers with OCR to generate computation-ready text [24]. Extracted sections comprise legal facts and legal considerations, representing the case's main narrative and judicial arguments, and are stored in structured format for subsequent merging with CSV data.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.8221

Extraction Validation

The validation process systematically reviews extracted text to ensure that all mandatory sections specifically legal facts and legal considerations are present, that text formatting and special characters are properly preserved, that document boundaries and section separators are correctly identified, and that OCR errors in scanned documents are corrected [24].

Dataset Characteristic Analysis

An empirical analysis was conducted to rule out trivial information leakage between source and target fields. Given that both fields originate from the same decision document, the mean Jaccard overlap between source (legal facts and considerations) and target (verdict statement) was computed for all 404 pairs. The mean Jaccard overlap is 0.102 ± 0.027, indicating that fewer than one in ten unique tokens are shared across the source–target pair on average, which is well below the conventional 0.2 threshold used to flag direct leakage. The target-to-source length ratio averages 2.11×, with 99.0% of instances exhibiting a longer target than source; this is an inherent characteristic of Indonesian *amar putusan*, which often restates defendant identity, article references, and sentencing details verbosely. This confirms that the task is genuine generation, not retrieval or copy-paste. Lexical overlap that does occur is concentrated on domain-specific performative tokens (e.g., "*MENGADILI*", "*Menyatakan terbukti*") that constitute the linguistic register of verdicts and cannot be avoided without sacrificing semantic fidelity.

XSum Dataset Construction

The dataset follows XSum principles mapping lengthy documents to concise abstractive summaries that focus on final results. The document column comprises combined legal facts and considerations from PDF extraction; the summary column comprises verdict statements from CSV data.

Table 1. Structure of Indonesian Legal XSum Dataset

Column	Description	Source	Length
id	Unique document identifier	PDF filename	404 documents
document (Source)	Legal facts and considerations from PDF	PDF extract	127 words (mean)
summary (Target)	Verdict statement (amar putusan)	CSV data	268 words (mean)

Source : (Research Results, 2026)

Verdict statements serve as reference summaries due to their concise, normative nature reflecting case conclusions, consistent with XSum's extreme-summarization objective [25].

Text Preprocessing

The preprocessing step removes excess whitespace, eliminates irrelevant special characters, and normalises text, while retaining punctuation and article numbering because they serve critical semantic functions in legal documents [26]. A minimal preprocessing approach is adopted because aggressive cleaning can eliminate important linguistic structures and reduce Transformer performance on domain text [27].

Data Splitting

The preprocessed dataset is divided into train (70%, n=282), validation (15%, n=61), and test (15%, n=61) subsets with a fixed random seed of 42. This ratio is standard in modern NLP experiments, balancing training volume with objective evaluation [28], [29].

Tokenization

Tokenisation uses each Transformer model's built-in tokenizer with a maximum input length of 1024 tokens and target length of 512 tokens (for T5/mT5/DistilBART/BART), accommodating long legal documents while maintaining efficiency through padding and truncation [30], [31], [32].

Model Architecture and Baselines

Four fine-tuned Transformers are evaluated against four extractive and four classical abstractive baselines. The choice of these specific models is motivated by four complementary strategies rather than a repetition of architectural definitions. T5 Base Indonesia is selected as an Indonesian-pretrained encoder–decoder that controls for monolingual pretraining on the target language [33]. mT5 Small is selected to probe whether multilingual pretraining alone without task-specific Indonesian data suffices in the legal register [34].

DistilBART CNN is selected as a compact news-summarization-pretrained model to measure cross-domain transfer from English news to Indonesian legal text at reduced computational cost [35], [36]. BART Large XSum is selected because its pretraining objective (extreme one-sentence summarization) most closely matches the structural profile of Indonesian verdict statements [37].

Table 2. Extractive Baseline Methods

Method	Type	Description
First Sentence	Extractive	Select first sentence of document
First 2 Sentences	Extractive	Select the first two sentences of the document
Lead 200 Characters	Extractive	Fixed-length character extraction
Legal Verdict Extract	Domain-specific Extractive	Regex-based extraction of verdict (*MENGADILI*) section

Source : (Research Results, 2026)

Table 3. Abstractive Baseline Methods

Method	Architecture	Training
Seq2Seq + Attention	LSTM Encoder–Decoder + Attention	Supervised
Pointer-Generator Network	LSTM + Copy Mechanism	Supervised
Neural Autoencoder	LSTM Autoencoder	Semi-supervised
Template-based Generation	Rule-based	No training

Source : (Research Results, 2026)

Table 4. Transformer Models

Model	Architecture	Parameters
T5 Base Indonesia	Encoder–Decoder (T5)	223M
mT5 Small	Encoder–Decoder (mT5)	300M
DistilBART CNN	Encoder–Decoder (Distilled BART)	306M
BART Large XSum	Encoder–Decoder (BART)	406M

Source : (Research Results, 2026)

Unlike prior research using Lead-3 [38], this study uses First-1, First-2, and Lead-200 to match the characteristic length of Indonesian verdict statements. The Legal Verdict baseline is a rule-based extractive approach that extracts the MENGADILI section [39]. For abstractive baselines, Seq2Seq+Attention captures source-summary alignment, Pointer-Generator Networks enable direct word copying from source documents [40], [41], the Neural Autoencoder learns compressed latent representations [42], and the Template-Based method produces structured summaries by mapping excerpts into fixed templates [43].

Fine-Tuning Configuration

Fine-tuning is performed using the Hugging Face Transformers Trainer framework on a single NVIDIA Tesla T4 GPU (16 GB VRAM) provided by Google Colab. Training uses mixed-precision (fp16) for BART-family models and full-precision with the Adafactor optimiser plus gradient checkpointing for mT5 Small, the latter chosen because mT5 is numerically unstable under fp16 (a known issue reported in the mT5 release notes). For T5 Base Indonesia, DistilBART CNN, and BART Large XSum, the AdamW optimiser is used. All models are trained for a maximum of 15 epochs with early stopping on validation loss (patience = 3). Per-model hyperparameters were tuned on the validation split and are reported in Table 5 together with the effective batch size (the product of per-device batch size and gradient accumulation steps) and the resulting parameter-to-sample ratio.

Table 5. Fine-Tuning Hyperparameter Configuration

Parameter	T5 Base Indo	mT5 Small	DistilBART CNN	BART Large XSum
Parameters	223M	300M	306M	406M
Learning Rate	5e-5	5e-5	3e-5	3e-5
Optimizer	AdamW	Adafactor	AdamW	AdamW
Per-device batch	2	1	4	2
Grad. accumulation	2	8	1	2
Effective batch	4	8	4	4
Precision	fp16	fp32	fp16	fp16
Grad. checkpoint	No	Yes	No	No
Max input tokens	1024	1024	1024	1024
Max target tokens	512	512	512	512
Max epochs	15	15	15	15
Early stop patience	3	3	3	3
Weight decay	0.01	0.01	0.01	0.01
Warmup steps	500	500	500	500

Source : (Research Results, 2026)

All training runs converged and triggered early stopping before the 15-epoch cap. Total wall-clock training time ranged from approximately 40 minutes (DistilBART CNN) to 2 hours 10 minutes (BART Large XSum) on the Tesla T4 (Google Colab), with peak GPU memory utilisation between 10 GB and 15 GB. Decoding at inference time uses beam search (num_beams = 4) with no-repeat-n-gram size = 3 and a repetition penalty of 1.5 across all Transformer models to mitigate verbatim loops; for mT5 Small we additionally force the BOS token to the pad token to suppress the sentinel-token artefact inherited from its span-corruption pretraining objective. The parameter-to-training-sample ratio (e.g., ~1.4 million parameters per training sample for BART Large XSum over 282 training examples) is high and is therefore controlled via weight decay (0.01), warmup, early stopping, and conservative learning rates, following best practice in low-resource fine-tuning [44].

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.8221

Evaluation Metrics

Evaluation uses ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 to measure lexical overlap and semantic similarity between predicted and reference summaries. ROUGE metrics quantify n-gram and longest-common-subsequence overlap [45]:

$$ROUGE - N = \frac{\text{Matching n-grams}}{\text{Total n-grams in reference}} \quad (1)$$

$$ROUGE - L = \frac{2 \times P \times R}{P + R} \quad (2)$$

Where P and R are derived from LCS measurements. While ROUGE ensures lexical coverage essential for legal documents, it may not capture deeper semantic nuances in abstractive summarization. BERTScore evaluates semantic similarity using contextual embeddings from BERT, enabling capture of meaning beyond simple n-gram overlap [46], [47]:

$$P_{\text{BERT}} = \frac{1}{|\hat{z}|} \sum_{\hat{z}_j \in \hat{z}} \max_{z_i \in z} \left(z_i^T \hat{z}_j \right) \quad (3)$$

$$R_{\text{BERT}} = \frac{1}{|z|} \sum_{z_i \in z} \max_{\hat{z}_j \in \hat{z}} \left(z_i^\top \hat{z}_j \right) \quad (4)$$

$$F1_{\text{BERT}} = \frac{2 \times P_{\text{BERT}} \times R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (5)$$

To complement point estimates and address concerns regarding the reliability of pairwise rankings on a 61-sample test set, we additionally report a paired bootstrap significance test with n = 10,000 resamples on per-sample ROUGE-L. Each comparison reports the mean difference, the 95% confidence interval, a two-sided p-value, and a significance flag at α = 0.05

RESULTS AND DISCUSSION

This section presents the evaluation results of fine-tuned Transformer models for Indonesian legal document summarization, benchmarked against extractive and classical abstractive baselines. Performance is assessed using ROUGE metrics and BERTScore across 404 Supreme Court human trafficking case documents, followed by qualitative analysis and discussion of practical implications.

Quantitative Results

Quantitative evaluation reveals three clear performance tiers. Among extractive baselines, First-2-Sentences achieves the highest BERTScore F1 of 60.24, but all extractive methods remain below ROUGE-1 = 11, confirming that direct sentence copying cannot reconstruct the verbose verdict register of Indonesian TPPO *amar putusan*. Among classical abstractive baselines, Seq2Seq+Attention is the strongest performer (ROUGE-1 = 27.31, BERTScore F1 = 61.81), outperforming Pointer-Generator, Template-based, and Neural Autoencoder; the Autoencoder effectively fails in this setting (ROUGE-2 ≈ 0.01, BERTScore F1 = 43.13), indicating an inability to produce coherent bigrams under the available training budget. Among fine-tuned Transformers, results vary substantially across metrics, reflecting differences in inference configuration and pretraining alignment. T5 Base Indonesia leads on ROUGE-L (39.49) and BERTScore F1 (74.82), while mT5 Small achieves the highest ROUGE-1 (44.97) and T5 the highest ROUGE-2 (31.80). DistilBART CNN (ROUGE-1 = 43.82) and BART Large XSum (ROUGE-1 = 42.78) remain competitive but are no longer the top performers across all metrics. All four fine-tuned Transformers substantially outperform all extractive and classical abstractive baselines.

Table 6. ROUGE and BERTScore Performance Comparison

Category	Model/Method	R1	R-2	R-L	BERT-F1
Fine-tuned	T5 Base Indonesia	44.25	**31.80**	**39.49**	**74.82**
Fine-tuned	mT5 Small	**44.97**	22.93	21.88	72.65
Fine-tuned	DistilBART CNN	43.82	25.03	33.06	71.92
Fine-tuned	BART Large XSum	42.78	25.63	33.25	72.10
Extractive	First Sentence	7.22	1.41	5.23	58.16
Extractive	First 2 Sentences	10.86	2.05	7.07	60.24
Extractive	Lead 200 Characters	6.09	1.17	4.89	58.36
Extractive	Legal Verdict Extract	7.22	1.41	5.23	58.16
Abstractive	Seq2Seq+Attention	27.31	15.71	22.53	61.81
Abstractive	Pointer-Generator	19.70	9.89	16.64	57.87
Abstractive	Neural Autoencoder	4.43	0.01	4.43	43.13
Abstractive	Template-based	14.06	8.15	12.83	65.15

Source : (Research Results, 2026)

.

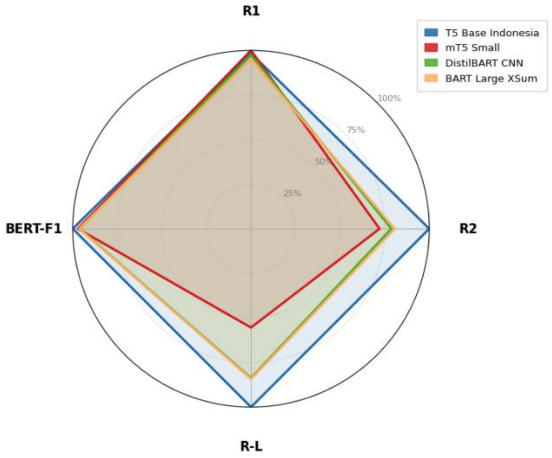
Statistical Significance of Pair-wise Differences

Table 7. Paired Bootstrap Significance Test

Model A vs Model B	Mean Δ(A–B)	95% CI	p-value	Significant (α=0.05)
T5 vs mT5	+17.6525	[+15.83, +19.44]	<0.001	Yes
T5 vs DistilBART	+6.4427	[+5.04, +7.78]	<0.001	Yes
T5 vs BART	+6.2580	[+4.82, +7.69]	<0.001	Yes
mT5 vs DistilBART	−11.2097	[−12.48, −9.97]	<0.001	Yes
mT5 vs BART	−11.3945	[−12.61, −10.22]	<0.001	Yes
DistilBART vs BART	−0.18	[−1.07, +0.74]	0.690	No

Source : (Research Results, 2026)

The paired bootstrap test on ROUGE-L confirms that T5 Base Indonesia leads all models: it outperforms DistilBART CNN (Mean Δ = +6.44, $p < 0.001$), BART Large XSum (Mean Δ = +6.26, $p < 0.001$), and mT5 Small (Mean Δ = +17.65, $p < 0.001$). DistilBART CNN and BART Large XSum remain statistically indistinguishable from each other (Mean Δ = −0.18, 95% CI [−1.07, +0.74], $p = 0.690$), and both robustly outperform mT5 Small ($p < 0.001$). Note that ROUGE-L rankings differ from ROUGE-1 rankings: mT5 Small leads on ROUGE-1 (44.97) but lags on ROUGE-L (21.88), reflecting its tendency to generate correct unigrams in non-sequential order. This ordering is stable under resampling and cannot be attributed to test-set noise. Figure 2 presents a radar chart contrasting the four fine-tuned Transformer models across ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 (normalized per metric), illustrating the lead of T5 Base Indonesia on ROUGE-L and BERTScore F1, and the divergence between ROUGE-1 and ROUGE-L performance for mT5 Small.



Source : (Research Results, 2026)

Figure 2. Radar Chart of Fine-Tuned Transformer Performance

Qualitative Analysis

Table 8. Sample Outputs *(Indonesian Language)* from Fine-Tuned Models

Model	Input/Document	Ground Truth/Summary	Model Prediction
T5 Base Indonesia	*Terdakwa menawarkan saksi korban (anak) kepada tamu melalui aplikasi MiChat di Hotel Grande Bandar Lampung pada 30 Mei Terdakwa mematok harga Rp500.000 untuk layanan seksual, di mana Terdakwa...*	*MENGADILI: Menyatakan Terdakwa Dila Aprilia Binti Suyadi tersebut di atas, terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana membantu melakukan perekrutan dengan...*	*MENGADILI: Menyatakan Terdakwa M. S. H. Als tersebut diatas, terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana Mengeksploitasi sebagaimana dakwaan alternatif pertama Penuntut Umum; Menjatuhkan...*
mT5 Small	*Bahwa Terdakwa Yesi Arisanti secara perseorangan melakukan perekrutan terhadap saksi korban (Zainal Abidin, dkk) untuk dipekerjakan di luar negeri (Jepang, Australia, Taiwan, dll) dengan meminta sejumlah uang...*	*MENGADILI: Menyatakan Terdakwa Yesi Arisanti Binti Bambang Marhaendi telah terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana Melakukan Penempatan Pekerja Migran Indonesia...*	*<extra_id_0> yang dijatuhkan; Menyatakan Terdakwa Yesi Arisanti secara sah dan meyakinkan bersalah melakukan penempatan Pekerja Migran Indonesia dengan pidana penjara selama 1 (satu) bulan; Menjatuhkan pidana kepada Terdakwa...*
DistilBART CNN	*Terdakwa menawarkan saksi korban (anak) kepada tamu melalui aplikasi MiChat di Hotel Grande Bandar Lampung pada 30 Mei Terdakwa mematok harga Rp500.000 untuk layanan seksual, di mana Terdakwa...*	*MENGADILI: Menyatakan Terdakwa Dila Aprilia Binti Suyadi tersebut di atas, terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana membantu melakukan perekrutan...*	*MENGADILI: Menyatakan Terdakwa M. SALEH BIN GITO tersebut diatas, terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana Turut serta melaksanakan perekrutan untuk tujuan mengeksploitasi orang...*
BART Large XSum	*Terdakwa Dzakiri An Nizami Bin Nur Cholis bersama saksi Muhammad*	*MENGADILI: Menyatakan Terdakwa Dzakiri An Nizami Bin*	*MENGADILI: Menyatakan Terdakwa Dzakiri An Nizami Bin Nur Cholis*

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.8221

Model	Input/Document	Ground Truth/Summary	Model Prediction
	Harun Alfatih telah mengeksploitasi anak korban Vivi Maulidiah (14 tahun) dengan cara menawarkan jasa...	*Nur Cholis tersebut diatas, telah terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana...*	*tersebut diatas, telah terbukti secara sah dan meyakinkan bersalah melakukan tindak pidana Eksploitasi Ekonomi dan Seksual...*

Source : (Research Results, 2026)

Table 9. Quantitative Error Analysis

Model	Hallucination	Repetition	Factual Error	Context Loss
T5 Base Indonesia	90.2%	4.9%	55.7%	14.8%
mT5 Small	88.5%	3.3%	31.1%	6.6%
DistilBART CNN	98.4%	13.1%	42.6%	0.0%
BART Large XSum	96.7%	0.0%	41.0%	1.6%

Source : (Research Results, 2026)

Subsequent to inference-time decoding optimisations specifically the application of n-gram repetition constraints and token-level penalty mechanisms for all models, with an additional decoder initialisation correction for mT5 Small the error profile shifts substantially. Repetition rates decline to near-zero for T5 Base Indonesia (4.9%) and mT5 Small (3.3%), confirming that the previously observed near-universal repetition was attributable to decoding configuration rather than underlying model incapacity. Nevertheless, hallucination of numeric entities persists at high rates across all four models (88.5–98.4%), indicating a fundamental generative limitation: models produce case identifiers, statutory article references, or sentencing durations that conform to the legal register but are unsupported by the source document. Factual error rates remain considerable across all models (31.1–55.7%), constituting the primary deployment risk verdict summaries containing erroneous numeric information cannot be reliably used for downstream legal retrieval or judicial reference without independent verification. Additionally, context loss is observed in 14.8% of T5 outputs and 6.6% of mT5 outputs, reflecting a completeness trade-off introduced by the repetition constraint; by contrast, DistilBART CNN and BART Large XSum maintain structural coverage (0.0% and 1.6% context loss respectively) at the expense of elevated hallucination rates.

Discussion

The quantitative and qualitative results carry four implications for Indonesian legal NLP. First, after applying decoding fixes, T5 Base Indonesia emerges as the strongest model on ROUGE-L (39.49) and BERTScore F1 (74.82), demonstrating that Indonesian-pretrained checkpoints with corrected inference yield strong performance on Indonesian legal text. mT5 Small leads on ROUGE-1 (44.97) but lags on ROUGE-L (21.88), suggesting it generates contextually relevant tokens without preserving their sequential structure a pattern that limits utility for legal documents where ordering is semantically critical. DistilBART CNN and BART Large XSum are statistically indistinguishable on ROUGE-L ($p = 0.690$) and provide a stable mid-range option. Second, classical Seq2Seq+Attention (ROUGE-1 = 27.31, BERTScore F1 = 61.81) remains a credible low-compute option for resource-constrained institutions. Third, hallucination of numeric entities (88.5–98.4% across all models) is the dominant deployment risk and motivates integration of constrained decoding or retrieval-augmented generation in follow-up work. Fourth, context loss in T5 (14.8%) and mT5 (6.6%) after repetition-penalty application suggests a trade-off between fluency control and coverage completeness that warrants further tuning.

From a deployment perspective, the artefacts of this study the XSum-formatted TPPO dataset, the n8n extraction workflow, and the fine-tuned BART Large XSum checkpoint are directly suitable for integration with Indonesian legal retrieval pipelines and case-management systems. Two concrete integration points are (i) Indonesia's e-court system (SIPP), where automatic summaries can be surfaced alongside the full decision to accelerate judicial triage, and (ii) public-facing legal databases (e.g., the Supreme Court directory), where automated summaries of TPPO decisions can lower the access barrier for victims' advocates, NGOs, and investigative journalists. Because hallucination remains a material risk, we recommend deploying these summaries as triage aids with prominent uncertainty indicators rather than as authoritative extracts.

Several limitations should be acknowledged explicitly. First, the corpus is restricted to 404 TPPO decisions from 2023–2024, which bounds claims of generalisation to other crime categories (e.g., narcotics, corruption) and to older jurisprudence. Second, reference summaries are derived solely from *amar putusan* and therefore omit the argumentation present in *pertimbangan hukum*; downstream users seeking reasoning-aware

summaries must combine these outputs with extractive retrieval over the considerations section. Third, comparison with other summarization approaches is literature-mediated via CivilSum and is therefore indicative rather than definitive; a direct evaluation using additional models on the present test set remains an open direction. Fourth, prior to the application of decoding optimisations, artefacts including sentinel-token leakage in mT5 Small and verbose repetition in T5 Base Indonesia were observed; while these were substantially mitigated, residual hallucination and context-loss patterns indicate that further decoding refinement or scaling to larger Indonesian-pretrained checkpoints warrants investigation.

CONCLUSION

This study delivered an XSum-structured Indonesian legal summarization dataset of 404 TPPO Supreme Court decisions, an n8n-based extraction pipeline, and a transparent benchmark of four Transformer architectures against four extractive and four classical abstractive baselines. The three contributions mapped onto the three research objectives are: (C1) an XSum-structured Indonesian TPPO dataset with explicit source–target separation and an empirically verified mean source–target Jaccard overlap of 0.102, (C2) the n8n extraction workflow that generalises across heterogeneous judicial PDFs, and (C3) a multi-tier benchmark accompanied by paired bootstrap significance testing and a quantitative error taxonomy.

The three main interpretive findings are: first, with decoding artefacts corrected, T5 Base Indonesia emerges as the strongest model on ROUGE-L (39.49) and BERTScore F1 (74.82), confirming the value of Indonesian-specific pretraining when inference is properly configured; mT5 Small leads on ROUGE-1 (44.97) but its lower ROUGE-L (21.88) reveals sequential-ordering weaknesses in the legal verdict register. Second, DistilBART CNN and BART Large XSum are statistically indistinguishable on ROUGE-L ($p = 0.690$) and offer stable cross-domain transfer from English news pretraining, while classical Seq2Seq+Attention remains a credible low-compute baseline (ROUGE-1 = 27.31). Third, hallucination of numeric entities persisting at 88.5-98.4% across all models even after decoding fixes is the dominant deployment risk and must be addressed before operational use.

Future research should pursue four concrete directions. (i) Expand the corpus beyond TPPO to narcotics, corruption, and domestic violence cases, and extend to first-instance and appellate decisions for broader generalisation. (ii) Evaluate additional summarization models directly on the present test set to close the literature-mediated comparison gap with published benchmarks. (iii) Integrate retrieval-augmented generation or constrained decoding to mitigate numeric-entity hallucination, which is the principal safety risk identified in the quantitative error analysis. (iv) Prototype integration with Indonesia's e-court system (SIPP) and the Supreme Court's public directory, with a user study of legal practitioners to validate utility and uncertainty communication in operational settings.

REFERENCE

[1] H. Abo-Bakr and S. A. Mohamed, "Automatic multi-documents text summarization by a large-scale sparse multi-objective optimization algorithm," *Complex & Intelligent Systems*, vol. 9, pp. 4629–4644, Feb. 2023, doi: 10.1007/s40747-023-00967-y.

[2] H. Y. Koh, J. Ju, M. Liu, S. Pan, and S. Pan, "An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics," *ACM Comput. Surv.*, vol. 55, no. 8, Aug. 2023, doi: 10.1145/3545176.

[3] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[4] M. Kirmani, G. Kaur, and M. Mohd, "Analysis of Abstractive and Extractive Summarization Methods," *International Journal of Emerging Technologies in Learning*, vol. 19, no. 1, pp. 86–96, 2024, doi: 10.3991/ijet.v19i01.46079.

[5] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences (Switzerland)*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137620.

[6] N. Sharma and S. Singh, "A Hybrid Extractive and Encoder-Decoder-Based Approach for Mitigating Hallucination in Automatic Text Summarization," *Journal of Transformative Technologies and Sustainable Development*, vol. 9, p. 15, 2025, doi: 10.1007/s41314-025-00083-4.

[7] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*,

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

vol. 55, no. 12, Dec. 2023, doi: 10.1145/3571730.

[8] Y. Song *et al.*, "A Hybrid Summarization Model for Legal Judgment Document Based on Domain Knowledge," *Information Technology and Control*, vol. 53, no. 3, pp. 772–784, 2024, doi: 10.5755/j01.itc.53.3.36602.

[9] A. Rao, S. Aithal, and S. Singh, "Single-Document Abstractive Text Summarization: A Systematic Literature Review," *ACM Comput. Surv.*, vol. 57, no. 3, Nov. 2024, doi: 10.1145/3700639.

[10] K. M. Rani Krishna, K. Somasundaram, P. Arulmozhivarman, S. A. Immanuel, and E. R. Rajkumar, "Deep learning for text summarization using NLP for automated news digest," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-20224-1.

[11] J. V. Alves, J. Liébana, H. Ferreira, and P. Bizarro, "SARSum: A Relevance and Comprehensiveness-Aware Abstractive Summarization Dataset for Suspicious Activity Reports," in *Frontiers in Artificial Intelligence and Applications*, IOS Press BV, Oct. 2025, pp. 4065–4072. doi: 10.3233/FAIA251296.

[12] M. Heddaya, K. MacMillan, H. Mei, C. Tan, and A. Malani, "CaseSumm: A Large-Scale Dataset for Long-Context Summarization from U.S. Supreme Court Opinions," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque: Association for Computational Linguistics, Apr. 2025, pp. 1917–1942. doi: 10.18653/v1/2025.findings-naacl.102.

[13] D. Datta, S. Soni, R. Mukherjee, and S. Ghosh, "MILDSum: A Novel Benchmark Dataset for Multilingual Summarization of Indian Legal Case Judgments," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 5291–5302. doi: 10.18653/v1/2023.emnlp-main.321.

[14] V. Zmiycharov, T. Tsonkov, and I. Koychev, "EurLexSummarization - A New Text Summarization Dataset on EU Legislation in 24 Languages with GPT Evaluation," in *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, Sofia: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sep. 2024, pp. 206–213. Accessed: Feb. 03, 2026. [Online]. Available: https://aclanthology.org/2024.clib-1.22/

[15] G. W. Wickasono, N. P. Hidayah, C. S. K. Aditya, A. F. P. Dewa, H. Fatikasari, M. A. P. Insani, M. H. F. Anwar, and N. A. Anhari, "Human Trafficking Court Decisions (Indonesia) — Structured Dataset," *Mendeley Data*, vol. 1, 2025, doi: 10.17632/8gtbky7r9x.1.

[16] Y. Sunusi, N. Omar, and L. Qadri Zakaria, "Exploring Abstractive Text Summarization: Methods, Dataset, Evaluation, and Emerging Challenges," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 1307–1324, 2024, doi: 10.14569/IJACSA.2024.01507130.

[17] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Natural Language Processing Journal*, vol. 7, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.

[18] K. Chaudhari, R. Mahale, F. Khan, S. Gaikwad, and K. Jadhav, "Comprehensive Survey of Abstractive Text Summarization Techniques," *International Research Journal on Advanced Engineering and Management*, vol. 2, no. 7, Jul. 2024, doi: 10.47392/IRJAEM.2024.0323.

[19] Y. Xue, X. Cao, X. Yang, Y. Wang, R. Wang, and J. Li, "We Need to Talk About Reproducibility in NLP Model Comparison," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 9424–9434. doi: 10.18653/v1/2023.emnlp-main.586.

[20] S. Casola, I. Lauriola, and A. Lavelli, "Pre-trained transformers: an empirical comparison," *Machine Learning with Applications*, vol. 9, p. 100334, Sep. 2022, doi: 10.1016/j.mlwa.2022.100334.

[21] A. Belz, S. Agarwal, A. Shimorina, and E. Reiter, "A Systematic Review of Reproducibility Research in Natural Language Processing," in *A Systematic Review of Reproducibility Research in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 381–393. doi: 10.18653/v1/2021.eacl-main.29.

[22] G. Hartawan, D. Sa'adillah Maylawati, and W. Uriawan, "Bidirectional and Auto-Regressive Transformer (BART) for Indonesian Abstractive Text

Summarization," *Jurnal Informatika Polinema*, vol. 10, no. 4, Aug. 2024, doi: 10.33795/jip.v10i4.5242.

[23] T. Hasan *et al.*, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 4693–4703. doi: 10.18653/v1/2021.findings-acl.413.

[24] S. D. Atagong, H. Tonnang, K. Senagi, M. Wamalwa, and K. M. Agboka, "A review on knowledge and information extraction from PDF documents and storage approaches," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1466092.

[25] H. Zhang, P. S. Yu, and J. Zhang, "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models," *ACM Comput. Surv.*, vol. 57, no. 11, Jun. 2025, doi: 10.1145/3731445.

[26] J. P. Tambusai, A. Maulana, N. Hafizhah Nst, N. Azahra, and Y. Lubis, "The Role of Punctuation In Clarifying Written Communication," *Jurnal Pendidikan Tambusai*, vol. 9, no. 1, Jan. 2025, doi: 10.31004/jptam.v9i1.25038.

[27] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, p. 102342, Mar. 2024, doi: 10.1016/j.is.2023.102342.

[28] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, May 2024, doi: 10.14569/IJACSA.2024.0150235.

[29] K. Hiniduma, S. Byna, and J. L. Bez, "Data Readiness for AI: A 360-Degree Survey," *ACM Comput. Surv.*, vol. 57, no. 9, Apr. 2025, doi: 10.1145/3722214.

[30] S. Yan, "Long Document Summarization with Transformer Models: A Comparative Evaluation on the Gutenberg Dataset," in *Proceedings of 2025 6th International Conference on Computer Information and Big Data Applications, CIBDA 2025*, Association for Computing Machinery, Inc, Aug. 2025, pp. 203–210. doi: 10.1145/3746709.3746746.

[31] Y. Sunusi, N. Omar, and L. Q. Zakaria, "Enhanced transformer for lengthcontrolled abstractive summarization based on summary output area," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.2667.

[32] R. Alsultan, A. Sagheer, H. Hamdoun, L. Alshamlan, and L. Alfadhli, "PEGASUS-XL with saliency-guided scoring and long-input encoding for multi-document abstractive summarization," *Sci. Rep.*, vol. 15, p. 26529, 2025, doi: 10.1038/s41598-025-11062-2.

[33] M. W. Bagus, D. Satya, and A. Luthfiarta, "Leveraging BERT and T5 for Comprehensive Text Summarization on Indonesian Articles," *Jurnal Sistem Cerdas*, vol. 8, no. 2, pp. 191–202, 2025, doi: 10.37396/jsc.v8i2.458.

[34] R. Dwi Putra, A. Rialdy Atmadja, and Y. A. Gerhana, "Improving Transformer Performance for Text Summarization in Video Transcription," *Journal of Computer Engineering, Electronics and Information Technology*, vol. 4, no. 2, pp. 83–90, 2025, doi: 10.17509/coelite.v4i2.89353.

[35] Gitanjali Mishra, Nilambar Sethi, Agilandeeswari Loganathan, Yu-Hsiu Lin, and Yu-Chen Hu, "Attention Free BIGBIRD Transformer for Long Document Text Summarization", *IJCISIM,* vol. 16, no. 2, p. 20, May 2024.

[36] H. Yuan and H. Zhang, "DomainSum: A Hierarchical Benchmark for Fine-Grained Domain Shift in Abstractive Text Summarization," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Association for Computational Linguistics, Apr. 2025, pp. 2219–2231. doi: 10.18653/v1/2025.findings-naacl.118.

[37] E. Daraghmi, L. Atwe, and A. Jaber, "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization Across Diverse Datasets," *Future Internet*, vol. 17, no. 9, Sep. 2025, doi: 10.3390/fi17090389.

[38] Y. Huang, L. Sun, C. Han, and J. Guo, "A High-Precision Two-Stage Legal Judgment Summarization," *Mathematics*, vol. 11, no. 6, p. 1320, 2023, doi: 10.3390/math11061320.

[39] D. Premasiri, T. Ranasinghe, R. Mitkov, M. El-Haj, and I. Frommholz, "Survey on legal information extraction: current status and open challenges," *Knowl. Inf. Syst.*, vol. 67, pp. 11287–11358, 2025, doi: 10.1007/s10115-025-02600-5.

[40] Mrunal Salwadkar, "Distributed Delivery Pipelines for Mobile and Ubiquitous Learning Platforms", *Secits Journal of Scalable Distributed Computing and Pipeline Automation,* vol. 1, no. 1, pp. 39–47, Dec. 2024, Accessed: May 25, 2026. [Online].

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.8221

Available: https://secitsociety.org/index.php/SJSDCPA/article/view/160

[41] A. Zhang, Z. C. Lipton, M. U. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge (atau New York/London sesuai imprint Cambridge): Cambridge University Press, 2024. Accessed: Feb. 02, 2026. [Online]. Available: https://www.cambridge.org/9781009389433

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. Accessed: Feb. 02, 2026. [Online]. Available: http://www.deeplearningbook.org

[43] S. Sharma, G. Aggarwal, and B. Kumar, "A survey on the dataset, techniques, and evaluation metric used for abstractive text summarization," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 22, no. 3, pp. 681–689, Jun. 2024, doi: 10.12928/TELKOMNIKA.v22i3.25512.

[44] N. Firdaus, B. Kusuma Riasti, and M. Asri Safi'ie, "Comparative Study Of Transformer-Based Models For Automated Resume Classification," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 11, no. 2, pp. 372–381, Nov. 2025, doi: 10.33480/jitk.v11i2.7453.

[45] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques," *J. Comput. Sci.*, vol. 87, May 2025, doi: 10.1016/j.jocs.2025.102571.

[46] Junadhi, Agustin, L. Efrizoni, F. Okmayura, D. R. Habibie, and Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 885–896, May 2025, doi: 10.47738/jads.v6i2.547.

[47] T. Sadikot and S. Jain, "Legal Document Summarization: A Zero-shot Modular Agentic Workflow Approach," in *Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*, Association for Computational Linguistics, 2025, pp. 29–40. Accessed: Feb. 02, 2026. [Online]. Available: https://aclanthology.org/2025.justnlp-main.5/