TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 24, No. 6, December 2026, pp. 2093~2102
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i6.28108

Financial table-to-text summarization of East Java government spreadsheets

Sofwatul Ummah, Supriyono, Okta Qomaruddin Aziz
Informatics Engineering Undergraduate Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia

Article Info

Article history:

Received Jun 11, 2026
Revised Sep 10, 2026
Accepted Oct 11, 2026

Keywords:

Abstractive summarization
Bidirectional and auto-regressive transformers
Factual consistency
Financial table-to-text
Government finance
TextRank

ABSTRACT

Government financial spreadsheets encode meaning through account labels, periods, units, and table structure, making source-faithful summarization difficult. This study evaluates an auditable table-to-text pipeline for East Java government spreadsheets, with source lineage, leakage-safe document splitting, fine-tuned bidirectional and auto-regressive transformers (BART), and a TextRank baseline. Eighteen Excel files yielded 266 documents and 816 token-bounded chunks. On 82 held-out chunks, TextRank achieved higher recall-oriented understudy for gisting evaluation (ROUGE)-1/2/L F1 than BART (0.7563/0.5929/0.7138 vs. 0.6321/0.4773/0.5917), although document-level differences were not significant after Holm correction. Researcher verification found factual inconsistencies in 71 of 82 BART summaries. In a blinded evaluation of 20 held-out documents by three independent evaluators, TextRank received significantly higher ratings for relevance, informativeness, fluency, and financial factual accuracy. The results show that auditable representation supports traceable evaluation, while abstractive generation still requires stronger factual safeguards before operational use in public-finance reporting.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Supriyono
Informatics Engineering Undergraduate Program, Faculty of Science and Technology
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Malang, Indonesia
Email: priyono@ti.uin-malang.ac.id

1. INTRODUCTION

Regional and provincial financial records are frequently distributed as spreadsheets in which fiscal meaning is carried by row labels, column roles, periods, units, worksheet context, and hierarchical account structure. A monetary value can therefore remain numerically plausible while becoming substantively wrong if it is attached to the wrong account. For example, assigning a Rp13.8 billion capital-expenditure value to operating expenditure would preserve a realistic number but change its accounting meaning and could mislead a reader about the composition of public spending. This study focuses on that relationship-preservation problem in East Java government financial spreadsheets.

Automatic text summarization has progressed from extractive selection toward neural abstractive and hybrid systems, with recent surveys emphasizing transfer learning, factuality, evaluation design, and domain dependence [1]–[5]. Pretrained encoder-decoder models such as bidirectional and auto-regressive transformers (BART) [6], pre-training with extracted gap-sentences for abstractive summarization (PEGASUS) [7], and text-to-text transfer transformer (T5) [8] made abstractive generation practical across many domains, but their performance depends on the language and data distribution used for pretraining and fine-tuning. For Indonesian, recent studies have explored multilingual BART (mBART) [9], multilingual T5

(mT5) [10], and multi-stage T5 transfer [11], while evidence from other low-resource languages also shows the value of language-specific pretrained representations [12].

Long and structured inputs introduce a second challenge. Surveys of long-document summarization identify context length, structural segmentation, and evaluation as persistent constraints [13]. Longformer/ longformer encoder-decoder (LED) [14] and BigBird [15] address long-sequence modeling through sparse-attention mechanisms, but using a different architecture would change the experimental question of this study. In parallel, table-to-text research increasingly emphasizes explicit structural context, fidelity, and copy accuracy [16]–[19]. These findings motivate preserving financial field relationships before a summarizer is applied, rather than treating spreadsheet flattening as a neutral preprocessing step.

The reliability requirement is especially important in financial natural language processing (NLP). Neural financial summarization has shown the potential of transformer-based generation [20], while broader finance research stresses traceability and explainability for artificial intelligence (AI)-assisted decision support [21], [22]. Factual-consistency studies likewise show that fluent summaries can contain unsupported or altered entities and relations [23], [24], and semantic or information-coverage metrics do not by themselves guarantee source-level financial correctness [25]. Accordingly, this work separates lexical overlap from source-grounded factual verification and independent human assessment.

The research gap is the lack of an auditable end-to-end procedure for transforming heterogeneous Indonesian government financial tables into traceable narrative facts under leakage-safe evaluation. BART was retained as the predeclared abstractive baseline because it provides a standard encoder-decoder framework, not because it is optimal for Indonesian financial language. Indonesian or multilingual domain adaptation, such as mBART/mT5 initialization or financial continued pretraining, is reserved for future work. The comparison of related studies and research positioning is summarized in Table 1.

Table 1. Comparison of related studies and research positioning

Study	Financial	Govt. data	Table-to-text	Abstractive	Long docs	Indonesian	Human eval
Lewis *et al*. [6]	✗	✗	✗	✓	✗	✗	✓
Zhang *et al.* [7]	✗	✗	✗	✓	✗	✗	✓
Astuti *et al*. [9]	✗	✗	✗	✓	✗	✓	✗
Beltagy *et al.* [14]	✗	✗	✗	✓	✓	✗	✗
Mo *et al*. [16]	✗	✗	✓	✓	✗	✗	✗
Yang *et al*. [18]	✗	✗	✓	✓	✗	✗	✗
Li *et al.* [20]	✓	✗	✗	✓	✗	✗	✓
This study	✓	✓	✓	✓	✓*	✓	✓

**Long-document handling in this study is implemented through source-preserving token-bounded chunking rather than a long-context model.*

This study evaluates an auditable summarization pipeline for East Java government financial spreadsheets. Its contributions include source-traceable preprocessing and serialization, leakage-safe no-truncation data splitting, and controlled BART–TextRank evaluation using automatic, factual, statistical, and human assessments. The contribution is the auditable representation and evaluation pipeline rather than a new model architecture.

2. METHOD

The framework consists of nine stages covering data collection, preprocessing and scale-aware normalization, source-traceable table-to-text transformation, document grouping and chunking, pseudo-label generation, leakage-safe splitting, BART fine-tuning, TextRank inference, and evaluation. Evaluation combines recall-oriented understudy for gisting evaluation (ROUGE), factual verification, paired statistical testing, and blinded human assessment, as illustrated in Figure 1.

2.1. Dataset collection

Eighteen publicly available East Java government Excel files were processed into 226 worksheets spanning seven financial document categories. The final corpus contained 266 document identities and 816 chunks, split at the document level into 212/27/27 documents for training, validation, and test (653/81/82 chunks). Leakage components kept shared facts, source values, pseudo-labels, and exact text within the same split; source-file overlap was diagnostic only, and the records contain institutional rather than personal data. The dataset statistics are summarized in Table 2.

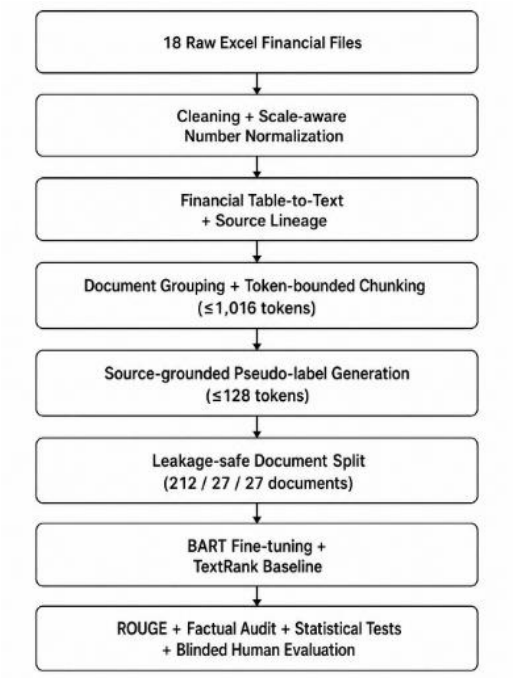


Figure 1. Auditable workflow for financial table-to-text summarization and multi-perspective evaluation

Table 2. Dataset statistics

Item	Value
Raw Excel files	18
Source worksheets	226
Document types	7
Independent documents	266
Final token-bounded chunks	816
Training set	212 docs; 653 chunks
Validation set	27 docs; 81 chunks
Test set	27 docs; 82 chunks
Input token statistics	Mean 786.02; max 1,016; no truncation

2.2. Data cleaning and structural normalization

Each worksheet was processed with persistent source identifiers linking records to their original file, worksheet, document, and financial values. Cleaning removed non-data fragments, normalized labels and whitespace, and preserved interpretable period, category, hierarchical account context, unit, and numeric context. Ambiguous rows were rejected rather than converted, enabling fail-closed source tracing for later factual and numeric verification.

2.3. Financial number normalization

Financial values were serialized into consistent Indonesian scale-aware forms using deterministic boundaries of (10^6) rupiah for juta and (10^9) rupiah for miliar, while preserving value lineage and numeric precision. A validation-only (3×3) sensitivity sweep perturbed these two boundaries to 0.9, 1.0, and 1.1 times their canonical values using the frozen BART model, without retraining, test-set access, or input truncation. All nine configurations remained within the 1,016-token input limit; canonical boundaries remained fixed, so the sweep measures input robustness rather than tuning.

2.4. Financial table-to-text transformation

Each accepted fact was converted into a self-contained Indonesian narrative preserving document context, period, category/account, value, sign, unit, and source lineage. The serializer binds categories to values, retains supported periods/comparisons and source order, stores document/worksheet/fact/value identifiers, and rejects unclear category-value relations. Example: *period 2023 + pembayaran sewa mobil + 146.241472 juta* becomes *pada periode 2023, pembayaran sewa mobil memiliki nilai keuangan sebesar 146,241472 juta rupiah.* The financial table-to-text serialization process is illustrated in Figure 2.

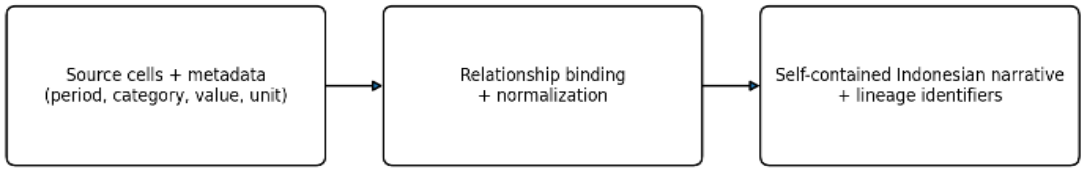


Figure 2. Financial table-to-text serialization: source fields are normalized and relationship-bound into an auditable narrative while source lineage is retained for downstream factual checking

2.5. Document grouping and token-bounded chunking

Source-grounded narratives were grouped by document and packed in source order into non-overlapping chunks capped at 1,016 BART tokens. This produced 816 chunks from 266 documents, with a tokenizer audit confirming that no source exceeded the limit in any split. Dynamic padding was therefore used during fine-tuning without source truncation, as summarized in Figure 3.

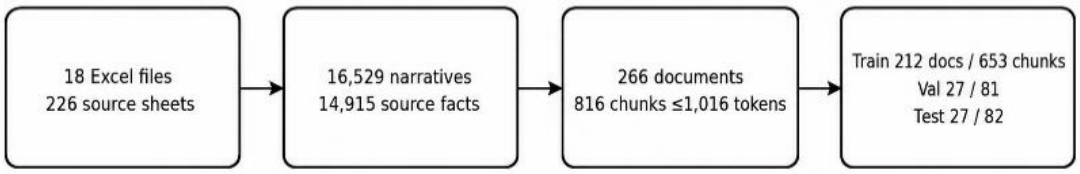


Figure 3. Dataset transformation, token-bounded chunking, and leakage-safe document split

2.6. Source-grounded pseudo-label generation

Because no human-authored references were available, deterministic source-grounded pseudo-labels were built within the 128-token budget, prioritizing distinct categories before repeats. These pseudo-labels are traceable but may favor lexical overlap and are not human ground truth. Therefore, ROUGE is interpreted only as similarity to the synthetic target and complemented by factual and blinded human evaluation.

2.7. BART fine-tuning

The Facebook/BART-base model [6] was fine-tuned on 653 training chunks from 212 documents, with 81 chunks from 27 documents used for validation only. Training used source sequences of at most 1,016 tokens and targets of at most 128 tokens with dynamic padding and no truncation. Overfitting was controlled through leakage-safe splitting, validation-only checkpoint selection, fixed five-epoch training, and complete exclusion of the test set from model decisions; no domain-adaptive pretraining or Indonesian-specific vocabulary expansion was performed, $L(\theta) = -\sum_{t=1}^{T} log\ P_\theta(y_t|y_{<t}, x)$.

Training used AdamW with learning rate 3×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, batch size 2, FP16, and seed 42 for five epochs, selecting the checkpoint with the lowest validation loss. Test inference used beam size 4, maximum/minimum lengths of 128/12, length penalty 1.0, no-repeat 3-gram constraint, and deterministic beam-search decoding, as summarized in Table 3.

Table 3. BART fine-tuning configuration

Parameter	Value
Base model	Facebook/BART-base (6 encoder / 6 decoder layers)
Input policy	≤1,016 source tokens; no truncation; dynamic padding
Max target length	128 tokens
Epochs	5; no training early stopping
Learning rate	3×10^{-5}
Batch size	2
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight decay	0.01
Checkpoint selection	Lowest validation loss
Inference	Beam = 4; min = 12; max = 128; no-repeat 3-gram
Hardware	NVIDIA Tesla T4; FP16

2.8. TextRank baseline

TextRank was retained as a training-free extractive control and was applied to exactly the same 82 held-out chunks from 27 documents as BART. It was selected not as a claim of state-of-the-art extractive performance, but because it isolates the effect of abstractive generation after both systems receive the same audited table-to-text representation [23]–[25]. Sentences were segmented with safeguards for common abbreviations, represented by lowercase unigram term frequency–inverse document frequency (TF-IDF) vectors, and connected by cosine similarity in a weighted graph. Weighted PageRank used damping 0.85 and tolerance 10^{-4}, ties were resolved by source position. Sentences were selected under the same 128-BART-token budget and restored to source order. More recent neural extractive [26] and hybrid pretrained [27] systems could provide stronger architecture-level baselines, but adding them after inspection of the held-out results would constitute a new experiment; they are therefore identified as future comparison work rather than retrofit to the locked test analysis.

2.9. Evaluation protocol

Primary evaluation used ROUGE-1/2/L against synthetic pseudo-labels at chunk and document-balanced levels. Exact two-sided Wilcoxon tests with Holm correction used 27 paired documents; all 82 BART outputs underwent factual verification. For 20 held-out documents, three independent evaluators (S2 Accounting/Government Accounting, S1 Public Administration/BPKAD, S2 Management/BPK Audit), each with 1–2 years of relevant Indonesian government-finance experience, rated relevance, informativeness, fluency, and financial accuracy on anchored 1–5 scales (1 = poor/severe, 3 = adequate/mixed, 5 = strong/fully supported), consented to anonymized reporting, and agreement used Fleiss' Kappa [28] and exact agreement. Paired human comparisons used Wilcoxon-Holm. A targeted ablation replaced only narrative table-to-text with field-value linearization while keeping other settings fixed.

Uncertainty was quantified using 95% percentile bootstrap confidence intervals (CIs) from 100,000 document-level resamples of the 27 held-out documents (seed 42). Supplementary post-hoc evaluation on the locked 82 test predictions used SacreBLEU [29] and BERTScore [30] with fixed configurations. These metrics were descriptive only and, because the references are synthetic pseudo-labels, were interpreted as complementary reference-similarity measures rather than tests of financial factual correctness.

2.10. Data and code availability

The original spreadsheets were obtained from publicly accessible East Java Provincial Government publications. Processed data, scripts, and audit artifacts are available from the corresponding author upon reasonable request, subject to source terms; no permanent public repository identifier is claimed. Human-evaluation records use anonymized evaluator identifiers and qualifications.

3. RESULTS AND DISCUSSION

Results are organized into five parts covering training dynamics, automatic evaluation and robustness analyses, representative cases, blinded human evaluation, and factual-error analysis with engineering implications and limitations. ROUGE is interpreted only as similarity to synthetic pseudo-labels, while factual reliability is assessed against source-grounded text and independent human ratings.

3.1. Training dynamics

Fine-tuning completed all five predeclared epochs, with validation loss decreasing from 0.2894 to 0.1912 and no late-epoch degradation. Epoch 5 was selected as the best checkpoint under the validation-loss criterion, with a final training loss of 0.2180. The test set was excluded from training and checkpoint selection, as illustrated in Figure 4.

3.2. ROUGE and supplementary similarity metrics

Table 4 shows that TextRank achieved higher chunk-macro ROUGE-1/2/L scores than BART (0.7563/0.5929/0.7138 vs. 0.6321/0.4773/0.5917), with the same ordering under document-balanced aggregation (0.7221/0.6067/0.6949 vs. 0.6604/0.5385/0.6200). However, none of the paired document-level differences remained significant after Holm correction (adjusted p = 0.1310, 0.1482, and 0.1156), so TextRank showed only a descriptive ROUGE advantage. The 95% bootstrap CIs for document-balanced means were 0.6115–0.7131/0.4723–0.6100/0.5653–0.6792 for BART and 0.6742–0.7705/0.5402–0.6738/0.6458–0.7441 for TextRank, while paired inference remained based on the Wilcoxon-Holm analysis.

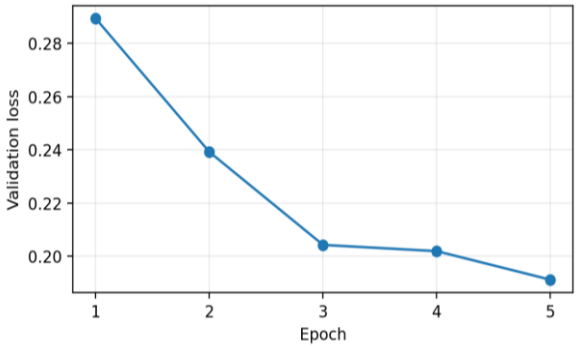


Figure 4. Validation loss during BART fine-tuning across five epochs

Table 4. Automatic evaluation metrics on the held-out test set (82 chunks / 27 documents)

Method	R-1 F1	R-2 F1	R-L F1	SacreBLEU	BERTScore F1 (chunk)	BERTScore F1 (doc.)
TextRank	0.7563	0.5929	0.7138	57.1559	0.9533	0.9467
Fine-tuned BART	0.6321	0.4773	0.5917	38.5953	0.9352	0.9354

ROUGE references were deterministic source-grounded pseudo-labels rather than human-written ground truth, so the scores reflect similarity to synthetic targets rather than factual accuracy. This distinction is important because overlap-based metrics may miss source-level entity and relation errors. A targeted retraining ablation replaced narrative inputs with deterministic field-value records while keeping facts, chunks, splits, targets, training, and generation settings fixed; Table 5 reports document-balanced ROUGE F1 on the same 27 held-out documents.

Table 5. Targeted table-to-text ablation on 27 held-out documents

Metric	Structured table-to-text	Field-value ablation	Holm p
ROUGE-1	0.6604	0.6739	0.9762
ROUGE-2	0.5385	0.5347	0.9762
ROUGE-L	0.6200	0.5974	0.9762

Field-value linearization was slightly higher on ROUGE-1 (0.6739 vs. 0.6604), while structured table-to-text was slightly higher on ROUGE-2/L (0.5385/0.6200 vs. 0.5347/0.5974); none of the paired differences was significant after Holm correction (adjusted $p = 0.9762$). A validation-only (3×3) boundary sweep with the frozen BART model produced only small ROUGE deviations, with maximum absolute document-balanced changes of 0.0020/0.0021/0.0004 for ROUGE-1/2/L, supporting low local sensitivity within the tested ±10% range rather than threshold optimization. Supplementary SacreBLEU and BERTScore results showed the same descriptive ordering in favor of TextRank, but were reported only as reference-similarity measures and not as factual-accuracy evidence or part of the primary inferential family.

3.3. Representative case analysis

Representative cases were selected after blinded human evaluation using predeclared deterministic rules rather than manual inspection. Table 6 shows the largest TextRank advantage (H19_ITEM_010), where BART preserves one valid asset value but introduces malformed wording, an unsupported location, and a damaged financial relation, whereas TextRank preserves three source sentences exactly. This illustrates the trade-off between abstractive compression and factual reliability in numerically constrained financial text, while extractive methods sacrifice paraphrasing and broader coverage. The result reflects the behavior of the two locked systems on this dataset and should not be interpreted as evidence that extractive summarization is universally superior.

3.4. Human evaluation

Three independent evaluators rated 40 blinded candidate summaries drawn from 20 distinct held-out documents. Agreement was high across all four dimensions: Fleiss' κ was 1.0000 for relevance, 0.9746 for informativeness, 1.0000 for fluency, and 0.8712 for financial factual accuracy; exact three-rater agreement

was 100.0%, 97.5%, 100.0%, and 87.5%, respectively. Table 7 reports mean system ratings. TextRank received higher ratings in every dimension. Paired item-level Wilcoxon tests remained significant after Holm correction for relevance ($p = 0.00098$), informativeness ($p = 0.00842$), fluency ($p = 0.000031$), and financial accuracy ($p = 0.000031$).

Table 6. Representative case from the blinded human evaluation (H19_ITEM_010)

Component	Example
Source context (excerpt)	*Dokumen Keuangan periode 2022 mencatat ASET sebesar 116,65427288808 miliar rupiah. Pada periode 2022, Akumulasi Amortisasi Aset Tidak Berwujud memiliki nilai keuangan sebesar -654,215689 juta rupiah. Pada periode 2022, JUMLAH EKUITAS memiliki nilai keuangan sebesar 116,25959987741 miliar rupiah. The full source contains no mention of "Surabaya".*
BART summary (overall mean = 1.00/5)	*Dokumen Keuangan periode 2022 mencatat ASET sebesar 116,65427288808 miliar rupiah. Lainnya memiliki nilai keuangan berbesar -654,215689 juta ruiah. Pada periodes 2022, JUMLAH EKUITAS/Defensit/Kurang-Kuituitan Pada Pada Ekumen Keunangan, Dokumen Ekumen dan Surabaya memilik nilai beruangan sebesars 116,25959987741 miliar rupiah.*
TextRank summary (overall mean = 4.0833/5)	*Dokumen Keuangan periode 2022 mencatat ASET sebesar 116,65427288808 miliar rupiah. Pada periode 2022, JUMLAH ASET memiliki nilai keuangan sebesar 116,65427288808 miliar rupiah. Pada periode 2022, JUMLAH KEWAJIBAN DAN EKUITAS memiliki nilai keuangan sebesar 116,65427288808 miliar rupiah.*

Table 7. Blinded human evaluation results on 20 held-out documents

Method	Relevance ($\kappa = 1.0000$)	Informativeness ($\kappa = 0.9746$)	Fluency ($\kappa = 1.0000$)	Fin. accuracy ($\kappa = 0.8712$)
TextRank	5.0000	2.8167	5.0000	4.9833
Fine-tuned BART	3.7500	2.1000	2.1000	2.4333

TextRank achieved a higher overall human-evaluation mean than BART (4.4500/5 vs. 2.5958/5), with the largest gaps in fluency (+2.90) and financial factual accuracy (+2.55), although these findings are limited to the 20-document blinded sample. The human and ROUGE results are complementary: ROUGE differences on 27 documents were not significant after Holm correction, whereas the blinded human study showed stronger differences favoring TextRank. This distinction is important in financial NLP, where fluent summaries must also preserve traceable categories, values, and relations [21], [22], [31], consistent with broader evidence that neural summaries may remain unreliable at entity and relation levels [23], [24].

3.5. Error analysis, engineering implications, and limitations

Researcher-conducted factual verification of all 82 BART test summaries found only 11 (13.41%) fully consistent with the locked source-grounded representation, while 71 (86.59%) contained at least one factual inconsistency. Unsupported additions were most frequent (64 summaries), followed by incorrect numeric values (50), value-period-category relation errors (36), period errors (28), damaged language (19), incomplete statements (13), category errors (12), and unit errors (1). The pattern suggests that free-form decoding can introduce unsupported content and alter bindings among closely related financial values, periods, and categories, although the observational audit does not establish causality. These findings motivate exact-value copying or whitelists, relation-aware validation, unsupported-token checks, and source-grounded human review. Persistent source identifiers, deterministic normalization, explicit value-period-category-unit bindings, and fail-closed validation support an auditable workflow from spreadsheet ingestion through summarization, automated checking, human approval, and downstream use in broader government information systems and traceable smart-governance workflows. Generated summaries should remain derivative artifacts rather than replacements for authoritative spreadsheets; no operational efficiency gain or production readiness is claimed because agency integration and field deployment were not evaluated.

Several limitations remain. The pseudo-labels are synthetic rather than human-authored, the systematic factual audit was researcher-conducted, and the dataset of 266 documents comes from a single provincial-government setting, limiting generalization. The fixed BART model was not adapted to Indonesian financial text, computational scalability was not benchmarked beyond the reported Tesla T4/FP16 setup, and no throughput, latency, memory-scaling, interoperability, or field-deployment evaluation was conducted. The retraining ablation was a fixed-protocol post-hoc comparison on the same 27-document test partition, whereas normalization-boundary sensitivity was restricted to validation data with a frozen model and local ±10% perturbations. ROUGE remains the primary inferential metric, supplementary similarity metrics are not factual-accuracy tests, and TextRank is a controlled training-free rather than modern neural baseline. Future work should therefore evaluate stronger neural baselines, factual and semantic metrics, human-authored references, broader normalization policies, Indonesian-domain pretraining, scalability, and independent multi-province replication and deployment.

4. CONCLUSION

This study developed and audited a financial table-to-text summarization pipeline for East Java government spreadsheets using leakage-safe document splitting, source-grounded pseudo-labels, and no-truncation tokenization. On 82 held-out chunks, TextRank achieved higher chunk-macro ROUGE-1/2/L F1 scores (0.7563/0.5929/0.7138) than BART (0.6321/0.4773/0.5917), although document-level differences were not significant after Holm correction. The table-to-text ablation likewise showed no significant effect, while a validation-only ±10% normalization-boundary sweep produced only small ROUGE deviations, indicating local robustness. Factual verification found only 11 of 82 BART summaries fully consistent with the source representation, with unsupported additions and incorrect numeric values as the dominant errors. In blinded human evaluation, TextRank also obtained substantially higher ratings than BART across relevance, informativeness, fluency, and financial factual accuracy.

These findings indicate that the evaluated BART configuration does not yet provide sufficient factual reliability for sensitive financial reporting, while source-preserving extractive methods remain a strong baseline when numerical and categorical fidelity are critical. The pipeline should therefore be viewed as an auditable, human-in-the-loop component in which source spreadsheets and lineage remain authoritative and automated factual checks act as release gates before generated summaries are used downstream. The pipeline has not been field-tested in a government agency; interoperability, security, scalability, user acceptance, and operational efficiency remain open questions beyond the controlled experiments. Future work should include broader multi-province datasets, human-authored references, Indonesian or multilingual financial-domain pretraining, stronger neural baselines, and constrained or fact-aware generation that explicitly preserves category-value-period relations.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Undergraduate Program in Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, for the academic support and facilities provided during this research.

FUNDING INFORMATION

This research did not receive any specific research grant or contract from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sofwatul Ummah	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Supriyono	✓	✓		✓	✓		✓			✓		✓	✓	
Okta Qomaruddin Aziz		✓		✓			✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

Informed consent was obtained from all three evaluators for participation in the study.

DATA AVAILABILITY

The original spreadsheets are publicly available through East Java Provincial Government publications. Processed data, scripts, and audit artifacts supporting the findings of this study are available from the corresponding author upon reasonable request, subject to the terms of the original data sources.

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, Mar. 2021, doi: 10.1016/j.eswa.2020.113679.

[2] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1029–1046, Apr. 2022, doi: 10.1016/j.jksuci.2020.05.006.

[3] A. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Computer Speech and Language*, vol. 71, p. 101276, Jan. 2022, doi: 10.1016/j.csl.2021.101276.

[4] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, vol. 603, p. 128255, Oct. 2024, doi: 10.1016/j.neucom.2024.128255.

[5] M. E. Saleh, Y. M. Wazery, and A. A. Ali, "A systematic literature review of deep learning-based text summarization: Techniques, input representation, training strategies, mechanisms, datasets, evaluation, and challenges," *Expert Systems with Applications*, vol. 252, p. 124153, Oct. 2024, doi: 10.1016/j.eswa.2024.124153.

[6] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.

[7] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," 2019, *arXiv*. doi: 10.48550/ARXIV.1912.08777.

[8] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2019, *arXiv*. doi: 10.48550/ARXIV.1910.10683.

[9] R. H. Astuti, M. Muljono, and S. Sutriawan, "Indonesian News Text Summarization Using MBART Algorithm," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 155–164, Feb. 2024, doi: 10.15294/sji.v11i1.49224.

[10] A. S. Wijaya and A. S. Girsang, "Augmented-Based Indonesian Abstractive Text Summarization using Pre-Trained Model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190–200, Nov. 2023, doi: 10.14445/22315381/IJETT-V71I11P220.

[11] S. Nasution, R. Ferdiana, and R. Hartanto, "Towards Two-Step Fine-Tuned Abstractive Summarization for Low-Resource Language Using Transformer T5," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, pp. 1220–1230, 2025, doi: 10.14569/IJACSA.2025.01602120.

[12] M. Bani-Almarjeh and M. B. Kurdy, "Arabic abstractive text summarization using RNN-based and transformer-based architectures," *Information Processing and Management*, vol. 60, no. 2, p. 103227, Mar. 2023, doi: 10.1016/j.ipm.2022.103227.

[13] H. Y. Koh, J. Ju, M. Liu, S. Pan, and S. Pan, "An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–35, Aug. 2023, doi: 10.1145/3545176.

[14] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," Dec. 2020, doi: 10.48550/arXiv.2004.05150.

[15] M. Zaheer *et al.*, "Big bird: Transformers for longer sequences," *Advances in Neural Information Processing Systems*, vol. 2020-December, Jan. 2020, doi: 10.48550/ARXIV.2007.14062.

[16] X. Mo, Y. Xiang, Y. Pan, Y. Hou, and P. Luo, "MGCOT: Multi-Grained Contextual Transformer for table-based text generation," *Expert Systems with Applications*, vol. 250, p. 123742, Sep. 2024, doi: 10.1016/j.eswa.2024.123742.

[17] I. Alonso and E. Agirre, "Automatic Logical Forms improve fidelity in Table-to-Text generation," *Expert Systems with Applications*, vol. 238, p. 121869, Mar. 2024, doi: 10.1016/j.eswa.2023.121869.

[18] Y. Yang, J. Cao, Y. Wen, and P. Zhang, "Table to text generation with accurate content copying," *Scientific Reports*, vol. 11, no. 1, p. 22750, Nov. 2021, doi: 10.1038/s41598-021-00813-6.

[19] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, and P. Gallinari, "Controlling hallucinations at word level in data-to-text generation," *Data Mining and Knowledge Discovery*, vol. 36, no. 1, pp. 318–354, Jan. 2022, doi: 10.1007/s10618-021-00801-4.

[20] H. Li, Q. Peng, X. Mou, Y. Wang, Z. Zeng, and M. F. Bashir, "Abstractive Financial News Summarization via Transformer-BiLSTM Encoder and Graph Attention-Based Decoder," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 3190–3205, 2023, doi: 10.1109/TASLP.2023.3304473.

[21] P. Weber, K. V. Carl, and O. Hinz, "Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature," *Management Review Quarterly*, vol. 74, no. 2, pp. 867–907, Jun. 2024, doi: 10.1007/s11301-023-00320-0.

[22] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 8, p. 216, Jul. 2024, doi: 10.1007/s10462-024-10854-8.

[23] Z. Luo, Q. Xie, and S. Ananiadou, "Factual consistency evaluation of summarization in the Era of large language models," *Expert Systems with Applications*, vol. 254, p. 124456, Nov. 2024, doi: 10.1016/j.eswa.2024.124456.

[24] Z. Chen and H. Lin, "Improving named entity correctness of abstractive summarization by generative negative sampling," *Computer Speech and Language*, vol. 81, p. 101504, Jun. 2023, doi: 10.1016/j.csl.2023.101504.

[25] D. M. Lal, K. P. Singh, and U. S. Tiwary, "ICE: Information coverage estimate for automatic evaluation abstractive summaries," *Expert Systems with Applications*, vol. 189, p. 116064, Mar. 2022, doi: 10.1016/j.eswa.2021.116064.

[26] Y. Wang, J. Zhang, Z. Yang, B. Wang, J. Jin, and Y. Liu, "Improving extractive summarization with semantic enhancement through topic-injection based BERT model," *Information Processing and Management*, vol. 61, no. 3, p. 103677, May 2024, doi: 10.1016/j.ipm.2024.103677.

[27] A. Ghadimi and H. Beigy, "Hybrid multi-document summarization using pre-trained language models," *Expert Systems with Applications*, vol. 192, p. 116292, Apr. 2022, doi: 10.1016/j.eswa.2021.116292.

[28] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa," *Behavior Research Methods*, vol. 57, no. 10, p. 287, Sep. 2025, doi: 10.3758/s13428-025-02746-8.

[29] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of*

the Conference, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 186–191. doi: 10.18653/v1/w18-6319.
[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2019, *arXiv*. doi: 10.48550/ARXIV.1904.09675.
[31] A. S. Bashir, A. A. Bichi, U. Mahmud, and A. M. Bello, "Long-Text Abstractive Summarization using Transformer Models: A Systematic Review," *Journal of the Brazilian Computer Society*, vol. 31, no. 1, pp. 1264–1279, Oct. 2025, doi: 10.5753/jbcs.2025.5786.

BIOGRAPHIES OF AUTHORS



Sofwatul Ummah is an undergraduate student in the Informatics Engineering Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia. Her research interests include natural language processing, automatic text summarization, machine learning, and financial document processing. She can be contacted at email: 230605110040@student.uin-malang.ac.id.



Supriyono is the Head of the Undergraduate Program in Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia. He received his undergraduate degree in Computer Science from Universitas Brawijaya in 2009 and his Master's degree in Informatics from Institut Teknologi Sepuluh Nopember Surabaya in 2012. He can be contacted at email: priyono@ti.uin-malang.ac.id.



Okta Qomaruddin Aziz is affiliated with the Informatics Engineering program at the Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia, and is associated with the Intelligent System Knowledge Group. He received his undergraduate degree in Computer Science from Universitas Indonesia in 2014 and his Master's degree in Computer Science from Universitas Indonesia in 2018. He can be contacted at email: okta.qomaruddin@uin-malang.ac.id.