

Trade-off Evaluation: BERT, RoBERTa, and DistilBERT for IMDb Sentiment Analysis on a Tesla T4 Google Colab Free Runtime

Asih Setiyorini^{*1}, Ema Utami²

¹ Informatika, PJJ Pascasarjana, Universitas Amikom Yogyakarta, 55283, Indonesia

² Program Doktor Informatika, Universitas Amikom Yogyakarta, 55283, Indonesia

Article Info

Article history:

Received June 23, 2026

Revised Aug 07, 2026

Accepted Sep 08, 2026

Keywords:

BERT

DistilBERT

RoBERTa

sentiment analysis

trade-off

ABSTRACT

This study addressed resource-constrained sentiment classification by balancing predictive quality with computational cost and stochastic variability. This study applied a five-stage structured evaluation protocol to compare BERT, RoBERTa, and DistilBERT on the 50,000-review IMDb dataset. The original experiment covered 27 configurations formed by three models, three random seeds, and three maximum sequence lengths. Architecture-level robustness was summarized using mean $\pm$ standard deviation across seeds, while one validation-selected configuration per architecture was retained only for confirmatory profiling on a uniform Tesla T4 Google Colab Free runtime. Performance was evaluated using accuracy, macro-F1, AUC-ROC, and MCC, and efficiency was assessed through checkpoint size, inference latency, and peak VRAM. Paired McNemar tests, bootstrap confidence intervals, classical TF-IDF baselines, Pareto analysis, and manual error inspection complemented the comparison. At sequence length 512, mean test macro-F1 was 0.95403 ± 0.00139 for RoBERTa, 0.94016 ± 0.00042 for BERT, and 0.93569 ± 0.00086 for DistilBERT. In the representative confirmatory runs, RoBERTa achieved the highest macro-F1 (0.9516), whereas DistilBERT provided the smallest checkpoint (256.10 MB), lowest BS16 latency (14.86 ms/sample), and lowest training VRAM (2399.80 MB). The results supported scenario-based model selection rather than a single universally optimal model.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Asih Setiyorini,

Informatika, PJJ Pascasarjana, Universitas Amikom Yogyakarta

Yogyakarta, 55283, Indonesia

Email: a.setiyorini@students.amikom.ac.id

1. INTRODUCTION

In sentiment classification, higher predictive scores do not necessarily indicate greater suitability for resource-constrained computing. Transformers capture textual context effectively (Jim et al., 2024; Mao et al., 2024; Minaee et al., 2022; Wankhade et al., 2022), but stronger predictive quality can require more memory and computation (Dehghani et al., 2022; Ganesh et al., 2021; Huang et al., 2022; Liu et al., 2024; Tay et al., 2023). Model comparison should therefore consider predictive quality and resource requirements together.

BERT, RoBERTa, and DistilBERT represent different Transformer design priorities. BERT is a bidirectional reference architecture (Devlin et al., 2019; Lin et al., 2022; Zhang & Shafiq, 2024), RoBERTa refines BERT pre-training (Liu et al., 2024), and DistilBERT uses knowledge distillation to reduce model size and computation (Sanh et al., 2019). Comparative studies show that rankings vary with datasets, hyperparameters, and evaluation conditions (Areshey & Mathkour, 2024; Bashiri & Naderi, 2024; Herawan & Saputri, 2025).

IMDb Movie Reviews was selected because it provides 50,000 balanced reviews and is widely used for sentiment classification (Puspita & Rahayu, 2023). Transformer-based methods have also performed strongly on IMDb (Islam et al., 2024; Paneru et al., 2025; Papia et al., 2024; Puspita & Rahayu, 2023). Fine-tuning is sensitive to stochastic factors, so repeated runs and seed-dependent optimization should be considered (Casola et al., 2022). Efficiency also depends on hardware, latency, memory, and input characteristics rather than parameter count alone (Ganesh et al., 2021; Tay et al., 2023).

Google Colab Free Tier was selected for accessibility, although GPU allocation and session continuity are not guaranteed. The confirmed hardware in this study was a Tesla T4 with 14.563 GB VRAM, so hardware-dependent results are interpreted only for the observed Tesla T4 Colab Free runtime. Rather than proposing a new architecture, this study applies a five-stage evaluation protocol linking stochastic variation, predictive performance, resource use, statistical evidence, error characteristics, and scenario-based recommendations.

This study asks: (1) how BERT, RoBERTa, and DistilBERT compare in predictive performance across tested seeds and in computational efficiency on the controlled Tesla T4 confirmatory runtime; (2) what performance-resource trade-offs emerge; and (3) what false-positive and false-negative characteristics are observed and how they inform model selection.

Prior benchmarks provide differently scoped evidence. (Islam et al., 2024) compared CNN, LSTM, CNN-LSTM, BERT+DNN, and RoBERTa+DNN on IMDB 50K using accuracy, precision, recall, and F1-score. (Areshey & Mathkour, 2024) compared five Transformers on Yelp under shared hyperparameters. (Cassee et al., 2024) used ten repeated runs with different random seeds and statistical testing, (Bashiri & Naderi, 2024) compared eight Transformers across 22 datasets, and (Herawan & Saputri, 2025) compared four Transformers on TeePublic reviews using accuracy, F1, and ANOVA. Together, these studies motivate a comparison that combines predictive robustness with controlled resource evidence.

Table 1. Systematic Positioning against Related Transformer Sentiment Studies

Study	Scope reported	Evaluation / reliability design	Primary comparison focus
(Areshey & Mathkour, 2024)	BERT, RoBERTa, ALBERT, DistilBERT, XLNet; Yelp reviews; shared hyperparameters	Comparative fine-tuning setup	Predictive performance
(Cassee et al., 2024))	2 Transformer and 3 ML sentiment tools; 2 software-engineering datasets	10 repeated train/test runs with different random seeds; Kruskal-Wallis and Dunn post-hoc tests	Predictive performance and meta-tokenization
(Bashiri & Naderi, 2024)	8 Transformer models across 22 sentiment datasets	Broad cross-dataset comparative analysis	Cross-dataset model performance
(Herawan & Saputri, 2025)	BERT, DistilBERT, RoBERTa, DeBERTa; 278,098 TeePublic reviews	Accuracy/F1 comparison with ANOVA	Predictive performance
(Islam et al., 2024)	CNN, LSTM, CNN-LSTM, BERT+DNN, RoBERTa+DNN; IMDB 50K reviews	Accuracy, precision, recall, and F1-score comparison	Predictive performance on IMDB
This study	BERT, RoBERTa, DistilBERT; IMDB; 3 seeds $\times$ 3 maximum lengths	Mean $\pm$ SD across seeds; exact McNemar; paired bootstrap; annotation reliability	Performance + T4 resource profiling + Pareto + error analysis

Table 1 shows that prior work addresses different parts of the evaluation problem. This study does not claim those elements are absent from earlier work. Its contribution is to integrate multi-seed robustness, controlled Tesla T4 profiling, paired statistics, Pareto interpretation, and error characterization within one resource-aware model-selection sequence.

2. METHOD

2.1. Structured Trade-off Evaluation Procedure

The study used a five-stage evaluation protocol rather than a new learning algorithm. Figure 1 preserves the original workflow: the 27-run grid, validation-based representative selection, uniform-environment confirmatory reruns, performance-efficiency evaluation, statistical and baseline validation, error and reliability analysis, and scenario-based integration. For the revised analysis, the 27 existing runs were also aggregated across seeds for each architecture-length setting (Table 4). The validation-selected confirmatory runs were retained only for Tesla T4 resource profiling, paired tests, Pareto analysis, and error inspection, preventing one representative seed from being treated as the architecture-level estimate.

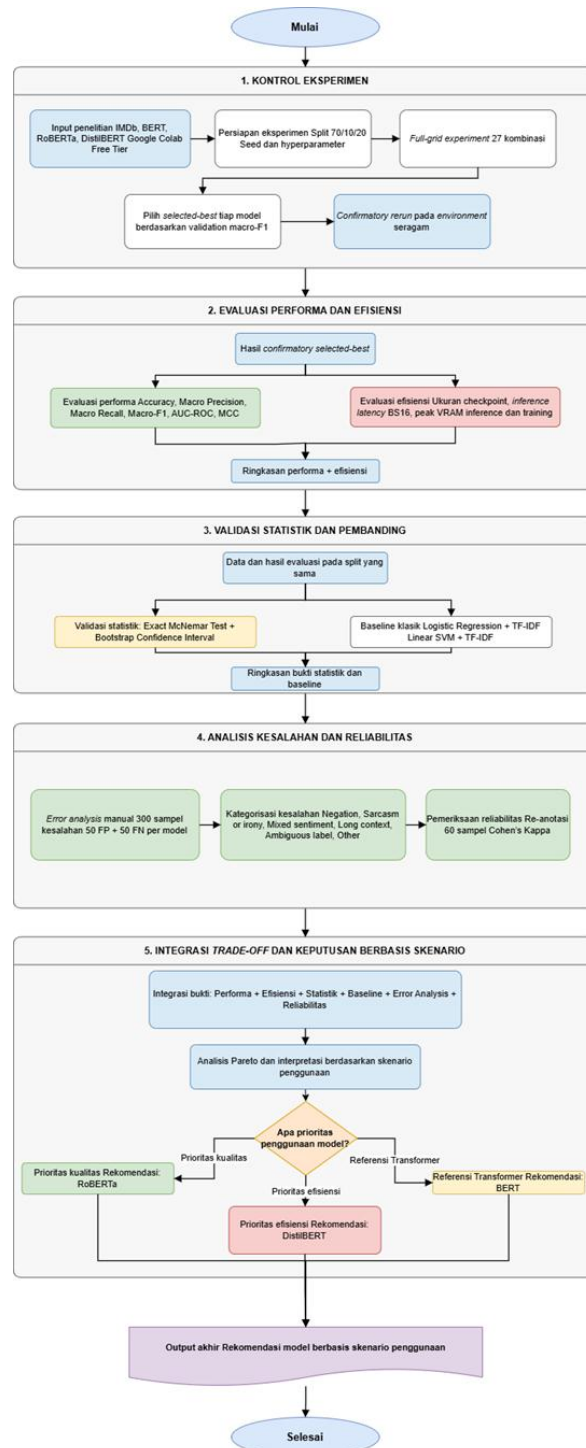


Figure 1. Five-Stage Performance-Efficiency Evaluation Protocol

2.2. Dataset and Experimental Design

The IMDb dataset contains 50,000 balanced English reviews, stratified into 35,000 training, 5,000 validation, and 10,000 test samples. Exploratory whitespace-token counts produced a mean of 231.16, median 173, 95th percentile 590, and 99th percentile 908; 3,694 reviews (7.39%) exceeded 512 tokens. The long-tailed length distribution supported testing maximum sequence lengths of 128, 256, and 512.

Whitespace-token counts were used only for exploratory analysis. During modeling, each architecture used its associated pre-trained tokenizer without stemming or lemmatization, and text beyond the specified maximum length was truncated. Three models, three seeds (42, 123, 2025), and three maximum lengths (128, 256, 512) produced 27 configurations. Table 2 summarizes the main grid settings.

Table 2. Main Experimental Configuration

Component	Value
Hugging Face checkpoints	bert-base-uncased; roberta-base; distilbert-base-uncased
Seed	42, 123, 2025
Maximum sequence length	128, 256, 512
Data split	70% training, 10% validation, 20% testing
Learning rate	2e-5
Train micro-batch / accumulation	8 / 2 (effective batch 16)
Maximum epoch / early stopping	5 / patience 2
Selection metric	Validation macro-F1

2.3. Model Configuration and Confirmatory Rerun

BERT, RoBERTa, and DistilBERT used bert-base-uncased, roberta-base, and distilbert-base-uncased. All models used learning rate 2e-5, weight decay 0.01, warmup ratio 0.10, evaluation batch size 16, maximum 5 epochs, and early stopping patience 2. The initial grid spanned Transformers 5.12.0, 5.12.1, and 5.13.1. Confirmatory reruns used Python 3.12.13, PyTorch 2.11.0+cu128, Transformers 5.13.1, Datasets 4.0.0, Accelerate 1.14.0, CUDA 12.8, Tesla T4 14.563 GB VRAM, and FP32. Training time is descriptive because sessions could be interrupted and resumed.

The original workflow selected one representative length-512 run per architecture by validation macro-F1: BERT seed 42, DistilBERT seed 2025, and RoBERTa seed 123. The revised interpretation treats seed as stochastic variation, so architecture-level results use mean $\pm$ SD across the three existing seeds for each model-length setting. Selected runs are used only for uniform Tesla T4 resource profiling, paired tests, Pareto analysis, and error inspection. Because the full grid spans several Transformers versions, its multi-seed summary is supporting robustness evidence rather than a fully version-controlled benchmark.

Table 3. Fine-tuning and Confirmatory Environment

Parameter	Value
Weight decay	0.01
Warmup ratio	0.10
Evaluation batch size	16
Maximum epoch	5
Early stopping patience	2
Python / PyTorch	3.12.13 / 2.11.0+cu128
Transformers / Datasets	5.13.1 / 4.0.0
Accelerate / CUDA	1.14.0 / 12.8
GPU / Precision	Tesla T4 14.563 GB / FP32

2.4. Performance, Efficiency, Statistical Validation, and Baseline

Predictive performance used accuracy, macro precision, macro recall, macro-F1, AUC-ROC, and Matthews Correlation Coefficient (MCC), with macro-F1 as the primary metric (Chicco et al., 2021; Chicco & Jurman, 2020). Full-grid mean $\pm$ SD across three seeds describes stochastic variation. Uniform confirmatory runs provide checkpoint size, batch-16 inference latency per sample, peak inference VRAM, and peak training VRAM. Training time is descriptive because confirmatory sessions were resumed after interruptions. Exact McNemar tests compare paired correct-incorrect outcomes on the common test set, while 1,000 paired-bootstrap iterations estimate 95% CIs for macro-F1 differences. These tests characterize the representative runs and complement the architecture-level multi-seed summaries.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Macro - F1 = \left(\frac{1}{K}\right) \sum_i \left[\frac{(2P_i R_i)}{(P_i + R_i)} \right] \quad (2)$$

In Equation (1), TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In Equation (2), P_i and R_i denote precision and recall for class i , and K =

2 represents the negative and positive sentiment classes. Accuracy is calculated as the proportion of all predictions that are correct, whereas macro-F1 gives equal weight to both classes.

Logistic Regression and Linear SVM with TF-IDF were evaluated on the same split for predictive-performance context. Both used 35,000 training and 10,000 test samples, with TF-IDF limited to 50,000 features, unigrams and bigrams, $\text{min_df} = 2$, and $\text{sublinear_tf} = \text{True}$. Latency was not compared directly across model families because their representations and execution devices were not equivalent; the baselines therefore provide classification context rather than cross-family efficiency claims.

2.5. Pareto, Error Analysis, and Reliability

Pareto analysis examined the trade-off among the three representative confirmatory configurations using macro-F1, checkpoint size, BS16 inference latency, and peak training VRAM without a weighted composite score. Error analysis then complemented the aggregate metrics by examining prediction-failure characteristics. From each model, 50 false positives and 50 false negatives were manually sampled, producing 300 annotated errors. Primary categories were Negation, Sarcasm or irony, Mixed sentiment, Long context, Ambiguous label, and Other.

Because the annotation was performed by a single annotator, the same 60 samples were annotated again using a design of 20 samples per model, consisting of 10 FP and 10 FN. Cohen's Kappa was used to assess intra-annotator reliability, while 2,000 bootstrap iterations were used to estimate the 95% confidence interval. The reliability result is interpreted as an indicator of categorical decision consistency rather than a substitute for inter-annotator reliability.

3. RESULT AND DISCUSSION

3.1. Multi-Seed Robustness and Confirmatory Performance-Efficiency

Across the full 27-run grid, maximum sequence length 512 produced the highest mean validation and test macro-F1 for all three architectures when results were aggregated across seeds 42, 123, and 2025. At length 512, mean test macro-F1 was 0.95403 ± 0.00139 for RoBERTa, 0.94016 ± 0.00042 for BERT, and 0.93569 ± 0.00086 for DistilBERT. The corresponding mean validation macro-F1 values were 0.95700 ± 0.00035 , 0.94373 ± 0.00153 , and 0.93827 ± 0.00231 . Thus, the architecture-level ordering at the longest tested input length remains RoBERTa > BERT > DistilBERT without relying on only the highest-scoring seed. Table 4 summarizes the mean and standard deviation for every model-length group. These aggregates come from the existing full-grid runs and are interpreted as robustness evidence because the initial grid spans Transformers versions 5.12.0, 5.12.1, and 5.13.1.

Table 4. Multi-Seed Macro-F1 Summary across the Full 27-Run Grid

Model	Max length	Validation macro-F1 mean $\pm$ SD	Test macro-F1 mean $\pm$ SD
BERT	128	0.89771 ± 0.00337	0.89267 ± 0.00114
BERT	256	0.92927 ± 0.00046	0.92449 ± 0.00123
BERT	512	0.94373 ± 0.00153	0.94016 ± 0.00042
DistilBERT	128	0.88766 ± 0.00065	0.88429 ± 0.00422
DistilBERT	256	0.91920 ± 0.00092	0.91609 ± 0.00227
DistilBERT	512	0.93827 ± 0.00231	0.93569 ± 0.00086
RoBERTa	128	0.91993 ± 0.00153	0.91242 ± 0.00189
RoBERTa	256	0.94233 ± 0.00061	0.93986 ± 0.00104
RoBERTa	512	0.95700 ± 0.00035	0.95403 ± 0.00139

For controlled hardware profiling, the original validation-selected representative configurations were retained: BERT_seed42_len512, DistilBERT_seed2025_len512, and RoBERTa_seed123_len512. These runs were repeated under the same Tesla T4 software environment and are used for Tables 5 and 6, the paired statistical tests, Pareto analysis, and error inspection. They are not interpreted as seed-averaged architecture estimates.

Table 5. Representative Confirmatory Classification Performance

Model	Accuracy	Macro-P	Macro-R	Macro-F1	AUC-ROC	MCC
RoBERTa	0.9516	0.9517	0.9516	0.9516	0.9884	0.9033
BERT	0.9415	0.9416	0.9415	0.9415	0.9790	0.8831
DistilBERT	0.9358	0.9358	0.9358	0.9358	0.9808	0.8716

Table 5 reports the representative confirmatory classification results. RoBERTa achieved the highest value across all reported classification metrics. BERT ranked above DistilBERT in accuracy, macro precision, macro recall, macro-F1, and MCC. However, the ordering differed for AUC-ROC: DistilBERT achieved 0.9808, slightly above BERT at 0.9790, while RoBERTa remained highest at 0.9884. Therefore, the relative ordering of BERT and DistilBERT is metric dependent rather than identical for every metric. Table 6 presents the computational efficiency profile of the same representative confirmatory runs, including checkpoint size, inference latency at batch size 16, peak inference VRAM, and peak training VRAM.

Table 6. Representative Confirmatory Efficiency Profile

Model	Val-F1	Test-F1	Checkpoint (MB)	Latency BS16 (ms)	Infer VRAM (MB)	Train VRAM (MB)
RoBERTa	0.9556	0.9516	478.91	30.30	2213.48	4572.17
BERT	0.9426	0.9415	418.35	34.47	1978.49	4334.99
DistilBERT	0.9392	0.9358	256.10	14.86	1344.79	2399.80

DistilBERT required fewer resources than RoBERTa and BERT in the representative Tesla T4 runs. Relative to RoBERTa, DistilBERT reduced inference latency by about 51.0%, checkpoint size by 46.5%, inference VRAM by 39.2%, and peak training VRAM by 47.5%. Its checkpoint was 256.10 MB versus 478.91 MB for RoBERTa, and latency was 14.86 versus 30.30 ms per sample. These percentages describe the representative confirmatory runs and are not seed-averaged hardware effects.

BERT occupied an intermediate position in terms of macro-F1 and checkpoint size, although its inference latency was slightly higher than that of RoBERTa. This finding indicates that model efficiency cannot be inferred directly from predictive performance alone. Instead, predictive quality, latency, storage requirements, and memory consumption should be evaluated jointly when selecting a model for a resource-constrained environment. Figure 2 further illustrates these performance-efficiency relationships across inference latency, checkpoint size, and peak training VRAM.

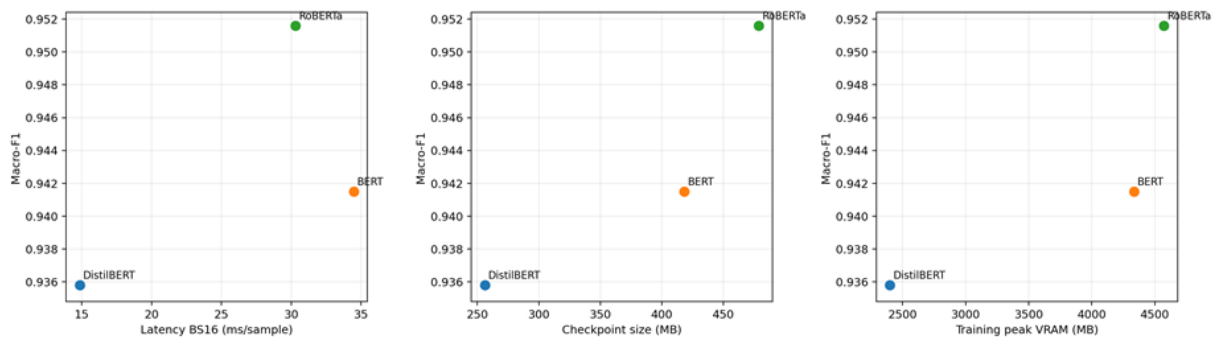


Figure 2. Pareto Analysis of Macro-F1 against Latency, Checkpoint Size, and Training VRAM

For the three representative confirmatory profiles, no single configuration dominated all evaluated dimensions. RoBERTa occupied the predictive-quality side, whereas DistilBERT occupied the efficiency side through lower latency, checkpoint size, and memory use. Because deployment priorities differ, no weighted composite score was used. The Pareto results therefore support requirement-driven selection rather than a universal winner.

3.2. Statistical Significance and Classical Baseline

For the representative confirmatory runs, exact McNemar tests gave $p = 0.00521$ for BERT-DistilBERT, $p = 6.74e-7$ for BERT-RoBERTa, and $p = 1.78e-12$ for DistilBERT-RoBERTa. Paired bootstrap 95% confidence intervals also excluded zero: [0.00181, 0.00940], [-0.01430, -0.00620], and [-0.02000, -0.01140], respectively. These tests characterize differences among the selected representative predictions and do not replace the multi-seed architecture-level summaries in Table 4.

Table 7. Pairwise Statistical Comparison of Representative Runs

Pair	Δ Macro-F1	McNemar p	Bootstrap 95% CI
BERT-DistilBERT	+0.00570	0.00521	[0.00181, 0.00940]
BERT-RoBERTa	-0.01010	6.74e-7	[-0.01430, -0.00620]
DistilBERT-RoBERTa	-0.01580	1.78e-12	[-0.02000, -0.01140]

The classical baselines provide context for the predictive performance of the three representative Transformer runs. Linear SVM with TF-IDF achieved a macro-F1 of 0.9138 and Logistic Regression achieved 0.9058, whereas the representative BERT, RoBERTa, and DistilBERT runs ranged from 0.9358 to 0.9516. Under the same data split, all three representative Transformer runs achieved higher macro-F1 than the two classical baselines. These results are used as performance context and not as evidence of cross-family efficiency superiority because the representation pipelines and execution devices differ. The classical baseline comparison is shown in Table 8.

Table 8. Classical Baseline and Representative Transformer Comparison

Model	Macro-F1	AUC-ROC	MCC
Logistic Regression TF-IDF	0.9058	0.9682	0.8117
Linear SVM TF-IDF	0.9138	0.9709	0.8276
DistilBERT	0.9358	0.9808	0.8716
BERT	0.9415	0.9790	0.8831
RoBERTa	0.9516	0.9884	0.9033

3.3. Error Analysis and FP-FN Pattern

After the quantitative trade-off between predictive performance and resource requirements was identified, error analysis was used to complement model selection with information not captured by aggregate metrics. Confirmatory predictions on the 10,000-review test set formed the basis for calculating false positives (FP) and false negatives (FN) for each model. Table 9 reports the false-positive and false-negative counts from the confirmatory predictions.

Table 9. Representative Confirmatory Prediction Errors

Model	FP	FN	Total errors	Error rate
RoBERTa	275	209	484	4.84%
BERT	330	255	585	5.85%
DistilBERT	338	304	642	6.42%

Table 10 summarizes the composition of the manually annotated subset of 300 prediction errors from the representative confirmatory runs.

Table 10. Primary Error Categories by FP and FN

Category	FP (n)	FP (%)	FN (n)	FN (%)
Mixed sentiment	52	34.67	85	56.67
Ambiguous label	63	42.00	20	13.33
Long context	15	10.00	16	10.67
Negation	5	3.33	14	9.33
Sarcasm or irony	12	8.00	5	3.33
Other	3	2.00	10	6.67

The FP-FN composition differed across categories. Ambiguous label was more common among false positives, whereas Mixed sentiment was more common among false negatives, consistently across all three models. Negation occurred more often among false negatives, while Sarcasm or irony, a recognized challenge in sentiment analysis (Băroiu & Trăușan-Matu, 2022), appeared more often among false positives. Because sampling was balanced at 50 FP and 50 FN per model, these results describe the annotated subset and are not prevalence estimates for the full test set.

Among 60 re-annotated samples, 32 retained the same primary category and 28 changed. Observed agreement was 53.33%, Cohen's Kappa was 0.3245, and the bootstrap 95% CI was [0.1308, 0.5022]. Disagreement clustered around Mixed sentiment, Ambiguous label, and Sarcasm or irony, so the categories are treated as supporting interpretive evidence rather than absolute labels. Table 11 reports the intra-annotator reliability results.

Table 11. Intra-annotator Reliability

N	Agreement	Disagreement	Observed	Kappa	Bootstrap 95% CI
60	32	28	53.33%	0.3245	[0.1308, 0.5022]

3.4. Comparison with Related Studies

The length-512 multi-seed ordering is consistent with the strong RoBERTa performance reported by (Areshey & Mathkour, 2024). Because prior benchmarks vary across datasets and protocols, architecture rankings should not be treated as universal. (Casseo et al., 2024) also illustrates the value of repeated runs and statistical testing. The present study adds controlled Tesla T4 profiling, paired prediction analysis, Pareto interpretation, classical baselines, and error characterization.

DistilBERT's lower checkpoint and VRAM requirements are consistent with the purpose of knowledge distillation (Ganesh et al., 2021; Sanh et al., 2019). RoBERTa achieved the highest predictive performance in these IMDb experiments, but this is not presented as a universal ranking. The contribution is the structured integration of complementary evidence for resource-aware selection under the observed conditions.

3.5. Methodological Contribution, Scenario-Based Recommendation, and Limitations

The methodological contribution is a structured, reusable evaluation protocol rather than a new architecture or algorithm. It combines experimental control, multi-seed robustness, representative hardware profiling, predictive and resource evaluation, paired statistics and baselines, error and reliability analysis, and Pareto-based scenario interpretation. Transferability beyond IMDb and the observed Tesla T4 runtime remains to be tested.

Limitations include English-language binary IMDb data, dynamic Colab Free conditions, Transformers-version variation across the initial grid, session-resumed training time, and single-annotator error analysis. The 27-run grid supports multi-seed robustness but is not a fully software-version-controlled hardware benchmark. The three Tesla T4 confirmatory reruns are validation-selected realizations rather than seed-averaged estimates. Accordingly, latency and memory are Tesla T4 observations; McNemar and bootstrap apply to the representative predictions; and error categories remain supporting interpretive evidence.

4. CONCLUSION

Joint interpretation of stochastic variation, predictive quality, and resource cost provided a more informative basis for model selection. At length 512, mean test macro-F1 across the existing seeds was 0.95403 ± 0.00139 for RoBERTa, 0.94016 ± 0.00042 for BERT, and 0.93569 ± 0.00086 for DistilBERT. Uniform Tesla T4 representative runs showed that DistilBERT required a smaller checkpoint, lower inference latency, and lower VRAM. Paired tests, Pareto analysis, classical baselines, and inspection of 300 errors added decision context, although Cohen's Kappa of 0.3245 limits error categories to supporting interpretation. Under the observed conditions, RoBERTa is favored for predictive quality, DistilBERT for resource efficiency, and BERT as a reference point. These recommendations are specific to IMDb and the observed Tesla T4 runtime.

ACKNOWLEDGMENTS

The authors thank the Master of Informatics Program and the Doctoral Program in Informatics, Universitas Amikom Yogyakarta, for academic support and facilities that supported this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding this research and publication.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Asih Setiyorini: Conceptualization; Data Curation; Methodology; Investigation; Formal Analysis; Visualization; Writing - Original Draft. Ema Utami: Conceptualization; Methodology; Supervision; Validation; Writing - Review & Editing.

REFERENCES

- Areshey, A., & Mathkour, H. (2024). Exploring transformer models for sentiment classification: A comparison of BERT, ROBERTA, ALBERT, DISTILBERT, and XLNET. *Expert Systems*, 41(11), e13701. <https://doi.org/10.1111/exsy.13701>
- Băroiu, A.-C., & Trăuşan-Matu, Ştefan. (2022). Automatic Sarcasm Detection: Systematic Literature Review. *Information*, 13(8), 399. <https://doi.org/10.3390/info13080399>
- Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowledge and Information Systems*, 66(12), 7305–7361. <https://doi.org/10.1007/s10115-024-02214-3>
- Casola, S., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: An empirical comparison. *Machine Learning with Applications*, 9, 100334. <https://doi.org/10.1016/j.mlwa.2022.100334>
- Cassee, N., Agaronian, A., Constantinou, E., Novielli, N., & Serebrenik, A. (2024). Transformers and meta-tokenization in sentiment analysis for software engineering. *Empirical Software Engineering*, 29(4), 77. <https://doi.org/10.1007/s10664-024-10468-2>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1), 13. <https://doi.org/10.1186/s13040-021-00244-z>
- Dehghani, M., Arnab, A., Beyer, L., Vaswani, A., & Tay, Y. (2022). *THE EFFICIENCY MISNOMER*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Ganesh, P., Chen, Y., Lou, X., Khan, M. A., Yang, Y., Sajjad, H., Nakov, P., Chen, D., & Winslett, M. (2021). Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. *Transactions of the Association for Computational Linguistics*, 9, 1061–1080. https://doi.org/10.1162/tacl_a_00413
- Herawan, D. F., & Saputri, T. R. D. (2025). Benchmarking Model Transformer Modern untuk Analisis Sentimen dan Tren Konsumen dalam Industri Fashion. *Edumatic: Jurnal Pendidikan Informatika*, 9(3), 945–954. <https://doi.org/10.29408/edumatic.v9i3.32657>
- Huang, S., Tang, E., Li, S., Ping, X., & Chen, R. (2022). Hardware-friendly compression and hardware acceleration for transformer: A survey. *Electronic Research Archive*, 30(10), 3755–3785. <https://doi.org/10.3934/era.2022192>
- Islam, M. T., Parvin, F., Sazan, S. A., & Bin Amir, T. (2024). Comparative Analysis of Sentiment Classification on IMDB 50k Movie Reviews: A Study Using CNN, LSTM, CNN-LSTM, and BERT Models. *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*, 512–517. <https://doi.org/10.1109/PEEIACON63629.2024.10800035>
- Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6, 100059. <https://doi.org/10.1016/j.nlp.2024.100059>
- Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, 3, 111–132. <https://doi.org/10.1016/j.aiopen.2022.10.001>
- Liu, H.-I., Galindo, M., Xie, H., Wong, L.-K., Shuai, H.-H., Li, Y.-H., & Cheng, W.-H. (2024). Lightweight Deep Learning for Resource-Constrained Environments: A Survey. *ACM Computing Surveys*, 56(10), 1–42. <https://doi.org/10.1145/3657282>

- Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4), 102048. <https://doi.org/10.1016/j.jksuci.2024.102048>
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep Learning--based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, 54(3), 1–40. <https://doi.org/10.1145/3439726>
- Paneru, B., Thapa, B., & Paneru, B. (2025). Sentiment analysis of movie reviews: A flask application using CNN with RoBERTa embeddings. *Systems and Soft Computing*, 7, 200192. <https://doi.org/10.1016/j.sasc.2025.200192>
- Papia, S. K., Khan, M. A., Habib, T., Rahman, M., & Islam, M. N. (2024). DistilRoBiLSTMFuse: An efficient hybrid deep learning approach for sentiment analysis. *PeerJ Computer Science*, 10, e2349. <https://doi.org/10.7717/peerj-cs.2349>
- Puspita, R., & Rahayu, C. (2023). Sentiment Analysis on IMDB Movie Reviews using BERT. *Indonesian Journal of Artificial Intelligence and Data Mining*, 6(2), 179. <https://doi.org/10.24014/ijaidm.v6i2.24239>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (Version 4). arXiv. <https://doi.org/10.48550/ARXIV.1910.01108>
- Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2023). Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6), 1–28. <https://doi.org/10.1145/3530811>
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780. <https://doi.org/10.1007/s10462-022-10144-1>
- Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, 11(1), 25. <https://doi.org/10.1186/s40537-023-00842-0>