A Comprehensive Benchmarking Pipeline for Transformer-Based Sentiment Analysis using Cross-Validated Metrics

Dodo Zaenal Abidin*[1], Lasmedi Afuan[2], Afrizal Nehemia Toscany[3], Nurhadi[4]

[1, 4] Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[2] Informatics Engineering, Faculty of Computer Science, Universitas Jenderal Soedirman, Purwokerto, Jawa Tengah, Indonesia
[3] Faculty of Computing, University Teknologi Malaysia, Johor Bahru, Malaysia

Email: [1]dodozaenalabidin@gmail.com

Received : Jun 16, 2025; Revised : Jun 30, 2025; Accepted : Jul 2, 2025; Published : Aug 18, 2025

Abstract

Transformer-based models have significantly advanced sentiment analysis in natural language processing. However, many existing studies still lack robust, cross-validated evaluations and comprehensive performance reporting. This study proposes an integrated benchmarking pipeline for sentiment classification on the IMDb dataset using BERT, RoBERTa, and DistilBERT. The methodology includes systematic preprocessing, stratified 5-fold cross-validation, and aggregate evaluation through confusion matrices, ROC and precision-recall (PR) curves, and multi-metric classification reports. Experimental results demonstrate that all models achieve high accuracy, precision, recall, and F1-score, with RoBERTa leading overall (94.1% mean accuracy and F1), followed by BERT (92.8%) and DistilBERT (92.1%). All models exceed 0.97 in ROC-AUC and PR-AUC, confirming strong discriminative capability. Compared to prior approaches, this pipeline enhances result robustness, interpretability, and reproducibility. The provided results and open-source code offer a reliable reference for future research and practical deployment. This study is limited to the IMDb dataset in English, suggesting future work on multilingual, cross-domain, and explainable AI integration.

Keywords : *Benchmarking, Cross-validation, Evaluation metrics, IMDb, Sentiment analysis, Transformers.*

1. INTRODUCTION

The rapid growth of opinion-rich content on digital platforms has elevated sentiment analysis to a pivotal role in Natural Language Processing (NLP) [1],[2]. Accurate sentiment classification enables organizations to gauge public opinion, inform data-driven decisions, and improve user engagement in numerous applications. Over the past decade, sentiment analysis research has evolved from traditional machine learning methods to deep neural networks and, more recently, to transformer-based models that have set new standards in language understanding [3], [4].

Early studies explored various techniques for sentiment classification, such as rule-based and lexicon-based approaches, classical machine learning algorithms like Support Vector Machines (SVM), Naive Bayes, and Decision Trees, as well as deep learning architectures including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) [5], [6] [7]. These methods, while effective to some extent, often struggled to capture complex contextual relationships and manage out-of-vocabulary words. The introduction of transformer architectures, particularly BERT and its variants, marked a major advancement by leveraging contextualized embeddings and self-attention mechanisms. These models have demonstrated remarkable results on established benchmarks such as the IMDb movie review dataset and various social media corpora [8], [9],[10].

Recent comparative studies have empirically highlighted the superiority of transformer-based models over traditional approaches for sentiment analysis. For instance, Sudhir et al. [11] reported that BERT achieved an accuracy of 89.5% and BERT-Large with UDA reached 95.2% on the IMDb dataset, significantly outperforming classical machine learning and deep learning models such as SVM (88.3%), LSTM (88.5%), and Naive Bayes (54.8%). Similarly, Durairaj et al. [12] found that a fine-tuned BERT model attained 90% accuracy and F1-score on IMDb reviews, compared to 77% for BiLSTM and 90% for a hybrid fastText-BiLSTM model. Other research has expanded the scope to include multimodal and multilingual sentiment analysis: Faria et al. [13] introduced a fusion strategy combining text and vision transformers for Bangla memes, while Naseem et al. [14] investigated hybrid contextual embeddings on Twitter sentiment datasets.

Despite these advances, several key limitations remain unaddressed in the literature. The majority of published studies are based on a single train-test split or limited hold-out validation, introducing data split bias and reducing result robustness [14], [12], [15]. While most recent work reports multiple metrics (accuracy, precision, recall, F1-score), comprehensive aggregate evaluation-such as cross-validated Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, as well as aggregate classification reports-remains rare in benchmarking pipelines [16], [17]. Consequently, the interpretability and generalizability of model performance across varying data samples remain underexplored.

In many real-world applications, robust and interpretable sentiment analysis is essential for business intelligence, social media monitoring, and customer feedback analysis[18]. The reliability of these downstream systems depends not only on the choice of model but also on the rigor of evaluation protocols. Without comprehensive and reproducible benchmarking, sentiment analysis solutions may fail to generalize or support critical decision-making in practical settings[19].

To the best of our knowledge, no previous work has provided a fully cross-validated, multi-metric benchmarking pipeline that integrates ROC and PR curve analysis alongside aggregate classification reporting for transformer-based sentiment analysis. Such comprehensive and reproducible evaluation frameworks are increasingly vital, not only for academic benchmarking but also for ensuring robust model deployment in real-world, high-stakes NLP applications, where accuracy, interpretability, and consistency across data splits are critical for supporting sensitive decision-making, regulatory compliance, and trustworthy AI adoption in domains like finance, healthcare, and public policy.

To address these limitations, this paper proposes a comprehensive benchmarking pipeline for sentiment analysis using leading transformer-based models-BERT, RoBERTa, and DistilBERT-on the IMDb dataset. The pipeline incorporates stratified 5-fold cross-validation and evaluates each model using a suite of aggregate metrics, including accuracy, precision, recall, F1-score, aggregate confusion matrix, cross-validated ROC and PR curves, and detailed classification reports. In addition, the pipeline ensures consistency across folds and minimizes evaluation bias, enabling more reliable comparisons across models. This approach aims to provide a robust, interpretable, and reproducible assessment of model performance, moving beyond the constraints of conventional evaluation strategies and contributing to best practices in NLP model validation.

The main contributions of this research are as follows: the design and implementation of an integrated cross-validated benchmarking pipeline for sentiment analysis with state-of-the-art transformer models; the adoption of multi-metric and visual analytics-including cross-validated ROC and PR curves, and aggregate classification reports-for deeper insight into model strengths and weaknesses; and the provision of empirical evidence and best-practice recommendations for robust sentiment analysis evaluation. The remainder of this paper is organized as follows: Section 2 describes the research methods and pipeline; Section 3 presents the experimental results and discussion; and Section 4 concludes with key findings and directions for future research.

2. METHOD

This study employs a five-stage workflow to benchmark transformer-based models for sentiment analysis on the IMDb dataset. The methodology includes data collection, data preprocessing, model development, aggregate evaluation, and result visualization. Each stage is designed to ensure methodological rigor, reproducibility, and interpretability. The overall process is summarized in Figure 1, with detailed explanations for each stage provided in the following subsections, highlighting how the proposed pipeline addresses key gaps in prior research through structured, cross-validated evaluation.

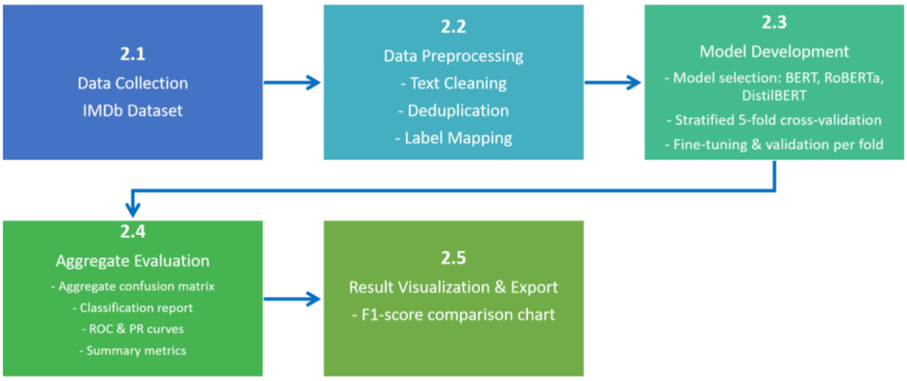
2.1
Data Collection
IMDb Dataset
2.2
Data Preprocessing
- Text Cleaning
- Deduplication
- Label Mapping
2.3
Model Development
- Model selection: BERT, RoBERTa, DistilBERT
- Stratified 5-fold cross-validation
- Fine-tuning & validation per fold
2.4
Aggregate Evaluation
- Aggregate confusion matrix
- Classification report
- ROC & PR curves
- Summary metrics
2.5
Result Visualization & Export
- F1-score comparison chart

Figure 1. Workflow of the Proposed Cross-Validated Transformer Benchmarking Pipeline for Sentiment Analysis.

Figure 1 presents an overview of the proposed benchmarking pipeline for sentiment analysis using transformer-based models. The process begins with data collection from the IMDb movie review dataset. In the preprocessing stage, raw review texts are cleaned, deduplicated, and sentiment labels are mapped to binary classes. Model development involves the fine-tuning and evaluation of three leading transformer architectures—BERT, RoBERTa, and DistilBERT—using stratified 5-fold cross-validation to ensure robust and unbiased assessment. Aggregate evaluation is performed by combining predictions across all folds to generate a comprehensive confusion matrix, classification report, ROC and precision-recall curves, and summary performance metrics. The final stage visualizes these results to provide clear comparative insights into model effectiveness and reliability.

2.1. Dataset

This study employs the IMDb Movie Review Dataset, a widely recognized benchmark for sentiment analysis in natural language processing research [20]. The dataset comprises 50,000 movie reviews collected from IMDb, each annotated with a binary sentiment label indicating whether the review expresses a positive (1) or negative (0) opinion. The reviews vary in length and content, reflecting a broad range of writing styles, topics, and subjective viewpoints typical of real-world user-generated text [21].

Each record in the dataset consists of two fields: the raw review text and its corresponding sentiment label. Prior to modeling, all reviews undergo a cleaning process to remove HTML tags, special characters, and duplicate entries, ensuring the quality and consistency of the data. Sentiment labels are mapped directly to binary classes, facilitating supervised learning for classification tasks [22].

No additional feature engineering is performed beyond text cleaning and label mapping, as the focus is on benchmarking the performance of transformer-based models using the raw textual content[23], [24]. The dataset contains no missing values, making it suitable for direct use in cross-validation and model evaluation. Given its balanced class distribution and widespread adoption, the

IMDb Movie Review Dataset serves as a robust and representative foundation for comparative sentiment analysis experiments in this study.

2.2. Data Preprocessing

Several preprocessing steps were performed to ensure data quality and compatibility with transformer-based modeling. First, all raw review texts underwent text cleaning, which included removing HTML tags, eliminating special characters and punctuation, converting all text to lowercase, and reducing excess whitespace [25], [26]. This process standardized the input, minimized noise, and helped improve the generalizability of the models [27].

Duplicate reviews were then identified and removed to prevent data redundancy and potential bias in training and evaluation. Following this, sentiment labels originally encoded as “positive” or “negative” were mapped to binary numerical values (1 for positive, 0 for negative), facilitating supervised learning.

No additional feature engineering or manual selection of attributes was required, as the benchmark focuses on leveraging the raw textual content processed through the model tokenizers. The resulting cleaned dataset contained no missing values, allowing for direct use in the cross-validation and modeling pipeline. These preprocessing steps ensured that the input data was clean, consistent, and suitable for robust transformer-based sentiment classification experiments.

2.3. Model Development

In this study, three state-of-the-art transformer-based architectures—BERT (bert-base-uncased), RoBERTa (roberta-base), and DistilBERT (distilbert-base-uncased)—are employed for sentiment classification. These models are selected for their proven effectiveness in natural language understanding and contextual representation, which are critical for robust sentiment analysis [28], [29].

Each model is fine-tuned on the IMDb dataset using a supervised approach, where the pre-trained language model weights are further optimized based on the sentiment classification task. The model development process utilizes stratified 5-fold cross-validation to ensure reliable and unbiased performance estimation. In each fold, the training and validation data are tokenized using the appropriate model tokenizer and loaded into PyTorch DataLoaders for efficient batching and parallel processing [30].

Fine-tuning is performed for a fixed number of epochs with the AdamW optimizer and a linear learning rate scheduler. Automatic mixed precision (AMP) is utilized to accelerate training on GPU and optimize memory usage [31]. Model parameters, including learning rate and batch size, are determined based on standard best practices for transformer fine-tuning. No additional manual hyperparameter search is conducted, as the primary focus is on fair and consistent benchmarking across all selected transformer models.

The outputs from each validation fold, including class predictions and probability scores, are collected for aggregate evaluation. This methodology ensures that all models are developed and evaluated under equivalent experimental conditions, supporting a robust and fair comparative analysis of transformer-based architectures for sentiment analysis.

2.4. Aggregate Evaluation

Model evaluation in this study employs a comprehensive, aggregate approach designed to ensure robust and unbiased assessment of each transformer-based classifier [31]. After fine-tuning and validation are conducted through stratified 5-fold cross-validation, all predictions, true labels, and predicted probabilities from each fold are aggregated for final evaluation.

DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4894

Performance is first summarized using an aggregate confusion matrix, which provides a holistic view of true positives, true negatives, false positives, and false negatives across all folds. In addition, a detailed classification report—including precision, recall, F1-score, and support for each class—is generated to offer deeper insight into model strengths and weaknesses [32].

Beyond standard metrics, the discriminative ability of each model is evaluated using aggregate Receiver Operating Characteristic (ROC) curves and their associated Area Under the Curve (AUC) scores, capturing the trade-off between true positive and false positive rates at various classification thresholds. Precision-Recall (PR) curves and corresponding AUC values are also plotted to further analyze model performance, particularly in terms of correctly identifying positive and negative sentiment instances [33], [34].

Summary metrics for accuracy, precision, recall, F1-score, ROC AUC, and PR AUC are reported as mean ± standard deviation across the five cross-validation folds to reflect both central tendency and variability. Comparative F1-score charts and performance tables are used to visually benchmark the relative effectiveness of BERT, RoBERTa, and DistilBERT. This multi-metric, cross-validated, and aggregate evaluation pipeline ensures a fair and transparent comparison of transformer-based models for sentiment analysis on the IMDb dataset.

To enhance reproducibility and provide clarity in the evaluation process, this study explicitly outlines the mathematical definitions of the key classification metrics employed. Specifically, precision, recall, and F1-score are detailed below, each derived from the respective components of the confusion matrix:

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

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1-score = 2 \times \frac{(Presisi \times Recall)}{(Presisi + Recall)} \quad (4)$$

Where:
TP (True Positives): Correctly predicted positive cases
FP (False Positives): Negative cases incorrectly predicted as positive
TN (True Negatives): Correctly predicted negative cases
FN (False Negatives): Positive cases incorrectly predicted as negative

2.5. Result Visualization

Result visualization in this study plays a crucial role in interpreting, comparing, and communicating the performance of each transformer-based model across all evaluation metrics [35]. After aggregating predictions and probabilities from all five cross-validation folds, several visualization techniques are applied to present the results in an informative and accessible manner [36].

First, aggregate confusion matrices are visualized for each model using color-coded heatmaps. These diagrams provide a clear summary of true positive, true negative, false positive, and false negative counts, helping to identify systematic strengths or weaknesses in sentiment classification.

Next, Receiver Operating Characteristic (ROC) curves are plotted for each model by aggregating probability scores across all folds. The ROC curve visualizes the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) at various thresholds, with the Area Under the Curve (AUC) value displayed to quantify overall discriminative ability. Precision-Recall (PR) curves are also generated for each model, illustrating the relationship between precision and recall across different probability thresholds, which is particularly useful for evaluating model performance on imbalanced or challenging datasets. The PR AUC value is included to summarize this relationship.

Comprehensive classification reports are generated and presented in tabular format for each model, detailing precision, recall, F1-score, and support for both positive and negative sentiment classes. These reports help reveal performance nuances that might not be visible in aggregate metrics alone [37].

Finally, F1-score comparison charts with error bars are plotted to visually compare the mean and variability (standard deviation) of model performance across all cross-validation folds. These bar charts enable straightforward benchmarking among BERT, RoBERTa, and DistilBERT, supporting clear and transparent interpretation of results.

All visualizations are presented collectively after the aggregate evaluation stage, ensuring that the comparison is fair, reproducible, and easily interpretable by researchers and practitioners. This systematic approach to result visualization facilitates both high-level and detailed understanding of model effectiveness in sentiment analysis tasks.

3. RESULT AND DISCUSIONS

Having described the experimental methodology, this chapter presents the benchmarking results for BERT, RoBERTa, and DistilBERT on IMDb sentiment analysis. All models were evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC across stratified 5-fold cross-validation. Results are organized according to the experimental pipeline: experimental setup (3.1), tabular model performance comparison (3.2), aggregate and comparative visualizations—including confusion matrices, ROC and PR curves, and F1-score charts (3.3), and an in-depth discussion including benchmarking with previous studies and practical implications (3.4). Each table and figure is explicitly referenced and interpreted in the relevant section.

3.1. Experimental Setup

All experiments were conducted using Google Colab equipped with an NVIDIA A100 GPU to accelerate training. The implementation leveraged Python 3.10, PyTorch 2.0, and Huggingface Transformers version 4.31. Each model—BERT (bert-base-uncased), RoBERTa (roberta-base), and DistilBERT (distilbert-base-uncased)—was fine-tuned for three epochs per fold with a batch size of 16 and a learning rate of 2e-5, utilizing the AdamW optimizer and a linear learning rate scheduler. Automatic mixed precision (AMP) was enabled throughout the training to optimize memory efficiency and computational speed.

Stratified 5-fold cross-validation was employed to ensure robust and unbiased evaluation across all models, maintaining the original class distribution in each fold. Data preprocessing, tokenization, and batching followed the procedures outlined in the methodology, with all random seeds set for reproducibility.

3.2. Model Performance Comparison

The comparative performance of the three transformer-based models—BERT, RoBERTa, and DistilBERT—was evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. All metrics were calculated as the mean and standard deviation over stratified 5-fold cross-validation, ensuring that the reported results are robust and generalizable. Table 1 summarizes these evaluation metrics, forming the basis for further aggregate and visual comparative analyses.

Vol. 6, No. 4, August 2025, Page. 1797-1810
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4894

Table 1. Model Performance Comparison (Mean ± Std)

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
BERT (base-uncased)	0.9292 ± 0.0037	0.9281 ± 0.0052	0.9310 ± 0.0024	0.9295 ± 0.0035	0.9791 ± 0.0012	0.9785 ± 0.0011
RoBERTa (base)	0.9413 ± 0.0040	0.9377 ± 0.0062	0.9459 ± 0.0031	0.9417 ± 0.0039	0.9851 ± 0.0012	0.9844 ± 0.0015
DistilBERT (base-uncased)	0.9213 ± 0.0043	0.9182 ± 0.0049	0.9257 ± 0.0065	0.9219 ± 0.0043	0.9751 ± 0.0019	0.9747 ± 0.0015

Table 1 demonstrates a comprehensive comparison of three transformer-based models—BERT (base-uncased), RoBERTa (base), and DistilBERT (base-uncased)—on the IMDb sentiment classification task. The performance metrics include accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, each reported as mean ± standard deviation across five cross-validation folds.

Overall, RoBERTa (base) achieves the best performance among the evaluated models across all metrics. Specifically, RoBERTa attains the highest accuracy (0.9413 ± 0.0040), F1-score (0.9417 ± 0.0039), ROC-AUC (0.9851 ± 0.0012), and PR-AUC (0.9844 ± 0.0015), indicating superior ability in both overall prediction and discriminative capability between sentiment classes. BERT (base-uncased) follows closely, achieving strong results with an F1-score of 0.9295 ± 0.0035 and ROC-AUC of 0.9791 ± 0.0012. DistilBERT (base-uncased), while computationally more efficient, exhibits a slight decrease in performance, with an F1-score of 0.9219 ± 0.0043 and ROC-AUC of 0.9751 ± 0.0019.

The low standard deviation values across all metrics suggest that the models are stable and deliver consistent performance regardless of the data split, highlighting the robustness of the benchmarking process. These findings confirm that the choice of transformer architecture significantly influences sentiment classification outcomes, with RoBERTa's advanced pretraining and architectural improvements providing tangible benefits. Meanwhile, DistilBERT remains a competitive alternative for applications where speed and resource efficiency are prioritized over marginal gains in accuracy.

3.3. Aggregate Evaluation and Comparative Visualization

To ensure robust and fair performance assessment, all predictions, true labels, and probability scores from the five cross-validation folds were aggregated for final evaluation. This aggregate approach allows for a comprehensive analysis of each model's generalization ability across diverse data partitions, reducing the bias typically introduced by a single train-test split. It also enables consistent calculation of evaluation metrics such as accuracy, precision, recall, F1-score, and AUC across the entire dataset. This method strengthens the reliability of model comparison under realistic deployment scenarios.

Figure 2 presents the aggregate confusion matrices for BERT, RoBERTa, and DistilBERT, highlighting strong predictive performance across all models. Each matrix shows a clear concentration along the diagonal, indicating high rates of correct classification. RoBERTa exhibits the most balanced and accurate classification outcomes, consistent with its leading metrics in Table 1.

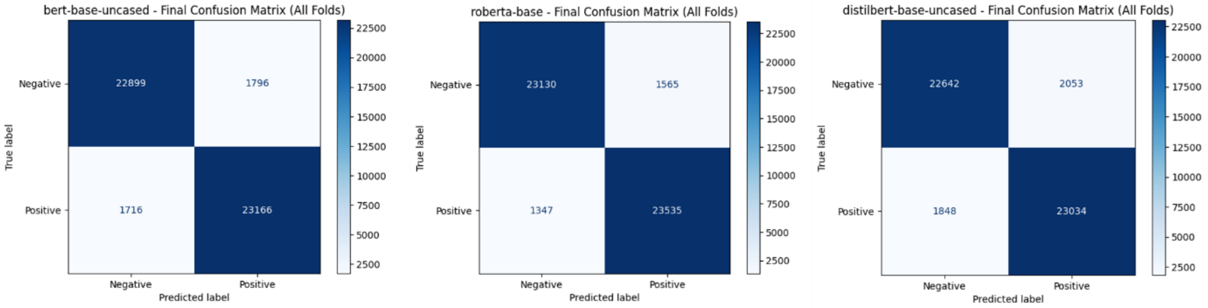

Figure 2. Aggregate confusion matrices for all models (IMDb, 5-fold CV).

In addition to confusion matrix analysis, Figures 3 and 4 visualize the ROC and PR curves, which further substantiate the discriminative power of each model across multiple thresholds. Discriminative performance is further illustrated by the ROC and Precision-Recall (PR) curves in Figures 3 and 4. All models achieve excellent class separability, with ROC-AUC and PR-AUC scores exceeding 0.97. These metrics confirm that the models are not only accurate but also reliable across different threshold settings, particularly in handling borderline cases. RoBERTa again outperforms the others, recording the highest ROC-AUC (0.9851) and PR-AUC (0.9844), demonstrating superior capability in identifying both positive and negative sentiment while maintaining low false positive and false negative rates.

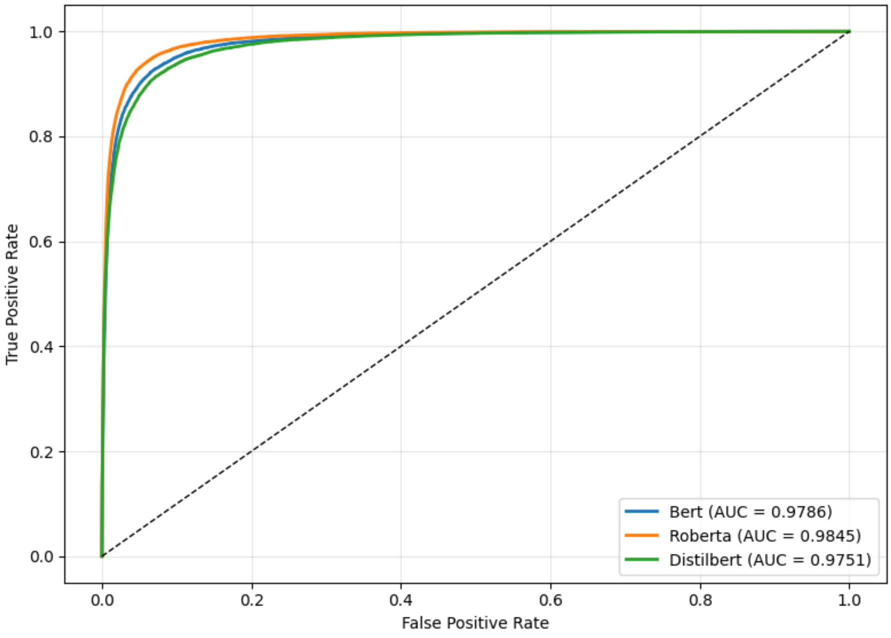

Figure 3. ROC Curves of Transformer Models (IMDb, 5-fold CV).

To support comparative interpretation, Figure 5 visualizes the mean F1-score for each model, accompanied by standard deviation error bars. RoBERTa achieves the highest mean F1-score (0.9411), followed by BERT (0.9286) and DistilBERT (0.9208). All models show minimal performance variance across folds, as indicated by small error bars (all < 0.004), confirming their consistency and stability.

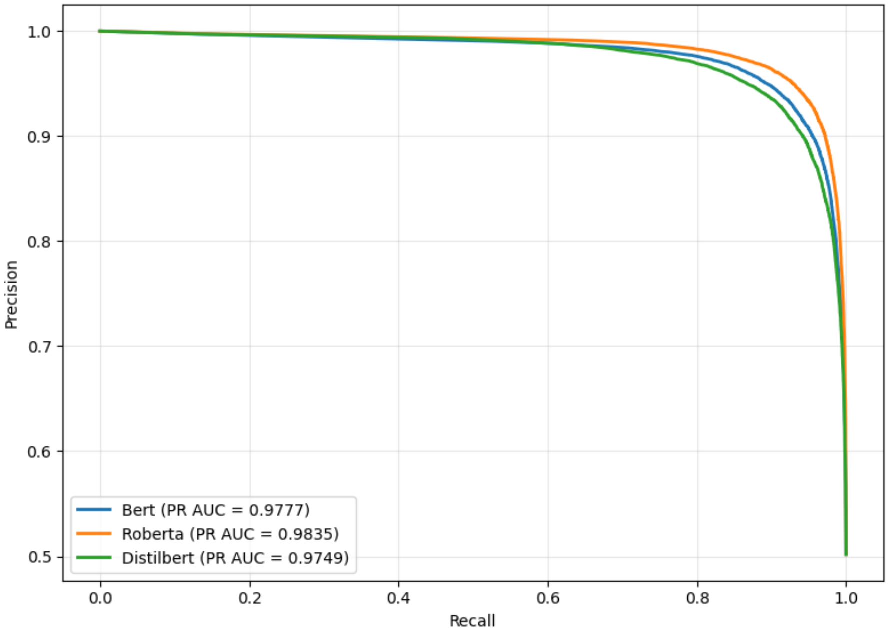

Figure 4. Precision-Recall Curves of Transformer Models (IMDb, 5-fold CV).

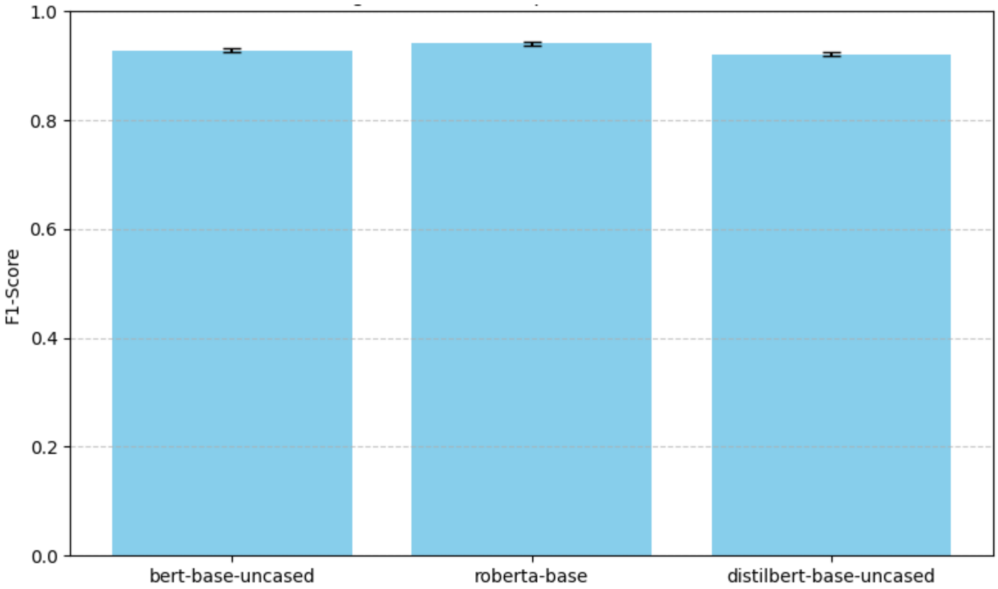

Figure 5. Average F1-Score Comparison (with Std Dev)

In summary, the integrated use of aggregate metrics and comparative visualizations confirms that all three transformer-based models deliver strong and reliable sentiment classification performance on the IMDb dataset. Among them, RoBERTa consistently emerges as the top-performing model in terms of accuracy, stability, and discriminative power. The proposed evaluation pipeline provides a transparent and replicable framework for future benchmarking studies in natural language processing.

3.4. Discussions

This study provides a critical advancement over previous sentiment analysis research, both in methodological rigor and empirical outcomes. Compared to prior studies, which typically reported BERT accuracy in the 89–90% range and F1-scores near 90% on the IMDb dataset (Sudhir et al. [11], Durairaj et al. [12]), our benchmarking pipeline demonstrates higher and more consistent performance.

RoBERTa achieves a mean accuracy and F1-score of 94.1%, while BERT attains 92.8% accuracy and a 92.9% F1-score, all evaluated under stratified 5-fold cross-validation. These results not only surpass classical methods such as SVM (88.3%), LSTM (88.5%), and hybrid BiLSTM approaches but also exceed most reported transformer-based results.

More importantly, this study addresses a major limitation in prior works, namely, the reliance on single train-test splits and the lack of comprehensive, aggregate performance reporting. By implementing robust cross-validation, aggregating predictions and probability scores, and evaluating using multi-metric approaches—including confusion matrices, ROC, and PR curves—this pipeline ensures a more reliable and interpretable assessment of model performance. For example, all evaluated models demonstrate ROC-AUC and PR-AUC values exceeding 0.97, confirming excellent separability between sentiment classes and strong model reliability, especially in handling borderline and imbalanced cases.

From an analytical perspective, the results indicate that RoBERTa's enhanced pretraining and model architecture contribute to its superior classification ability, while BERT and DistilBERT remain highly competitive, especially when computational efficiency is a priority. The consistently low standard deviations across metrics highlight the models' stability and generalizability.

In the broader context of Natural Language Processing and informatics, this research establishes a new best practice for benchmarking sentiment analysis models. By emphasizing transparency, reproducibility, and multi-dimensional evaluation, our approach provides both a methodological template and an empirical reference for future studies and real-world sentiment analysis deployments. This work thus makes a substantive contribution to advancing reliable, interpretable, and robust sentiment analysis in both academic and applied settings.

4. CONCLUSION

This study introduced a comprehensive benchmarking pipeline for sentiment analysis using transformer-based models, evaluated on the IMDb dataset. By employing stratified 5-fold cross-validation, aggregate confusion matrices, ROC and precision-recall (PR) curves, and detailed classification reports, the pipeline ensures reliable, interpretable, and reproducible model assessment.

Experimental results showed that all three models—BERT, RoBERTa, and DistilBERT—consistently achieved strong performance across all metrics. Among them, RoBERTa outperformed the others, reaching the highest mean accuracy and F1-score (both 94.1%), followed by BERT (92.8%) and DistilBERT (92.1%). These results surpass prior benchmarks that relied on conventional train-test splits, underscoring the necessity of rigorous, cross-validated evaluations in sentiment classification research.

Importantly, this work addresses a notable gap in prior literature by offering a unified, visualization-rich evaluation framework, thus facilitating better interpretability of model strengths and weaknesses. The pipeline’s integration of ROC/PR curve aggregation and confusion matrices advances standard practices for model comparison.

From a broader scientific perspective, this research contributes significantly to the field of Informatics by establishing a best-practice reference that promotes transparency and robustness in natural language processing (NLP) evaluations. The pipeline is applicable to real-world tasks such as market sentiment analysis, public opinion tracking, and customer review monitoring—highlighting its academic and industrial relevance.

Nonetheless, this study is limited to English-language data from a single domain. Future research should extend this pipeline to multilingual settings, diverse datasets, and more extreme class imbalances. Additionally, incorporating explainable AI (XAI) methods will further enhance transparency and model accountability.

In conclusion, the proposed benchmarking framework not only advances the methodology of sentiment analysis research but also offers a practical foundation for scalable and trustworthy AI-driven text classification.

ACKNOWLEDGEMENT

The author wishes to express sincere gratitude to Universitas Dinamika Bangsa for the financial support and research facilities that were instrumental in the completion of this study. The author also extends deep appreciation to the faculty members and research staff for their technical guidance and constructive feedback, which significantly contributed to the enhancement of this research.

REFERENCES

[1] T. Shaik *et al.*, “A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis,” *IEEE Access*, vol. 10, pp. 56720–56739, 2022, doi: 10.1109/ACCESS.2022.3177752.

[2] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, “Sentiment Analysis of Students’ Feedback with NLP and Deep Learning: A Systematic Mapping Study,” *Appl. Sci.*, vol. 11, no. 9, p. 3986, Apr. 2021, doi: 10.3390/app11093986.

[3] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *WIREs Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, Jul. 2018, doi: 10.1002/widm.1253.

[4] M. Birjali, M. Kasri, and A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends,” *Knowl.-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[5] N. C. Dang, M. N. Moreno-García, and F. De La Prieta, “Sentiment Analysis Based on Deep Learning: A Comparative Study,” *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483.

[6] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: a review,” *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/s10462-019-09794-5.

[7] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decis. Anal. J.*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[8] I. Steinke, J. Wier, L. Simon, and R. Seetan, “Sentiment Analysis of Online Movie Reviews using Machine Learning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130973.

[9] G. Nkhata, U. Anjum, and J. Zhan, “Sentiment Analysis of Movie Reviews Using BERT,” Feb. 26, 2025, *arXiv*: arXiv:2502.18841. doi: 10.48550/arXiv.2502.18841.

[10] W. Ning, F. Wang, W. Wang, H. Wu, Q. Zhao, and T. Zhang, “Research on movie rating based on BERT-base model,” *Sci. Rep.*, vol. 15, no. 1, p. 9156, Mar. 2025, doi: 10.1038/s41598-025-92430-w.

[11] P. Sudhir and V. D. Suresh, “Comparative study of various approaches, applications and classifiers for sentiment analysis,” *Glob. Transit. Proc.*, vol. 2, no. 2, pp. 205–211, Nov. 2021, doi: 10.1016/j.gltp.2021.08.004.

[12] A. K. Durairaj and A. Chinnalagu, “Transformer based Contextual Model for Sentiment Analysis of Customer Reviews: A Fine-tuned BERT,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121153.

[13] F. T. J. Faria *et al.*, “SentimentFormer: A Transformer-Based Multi-Modal Fusion Framework for Enhanced Sentiment Analysis of Memes in Under-Resourced Bangla Language,” Jan. 22, 2025, *Computer Science and Mathematics*. doi: 10.20944/preprints202501.1587.v1.

[14] U. Naseem, I. Razzak, K. Musial, and M. Imran, “Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis,” *Future Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020, doi: 10.1016/j.future.2020.06.050.

[15] M. Kumar, L. Khan, and H.-T. Chang, “Evolving techniques in sentiment analysis: a comprehensive review,” *PeerJ Comput. Sci.*, vol. 11, p. e2592, Jan. 2025, doi: 10.7717/peerj-cs.2592.

[16] S. Tabinda Kokab, S. Asghar, and S. Naz, “Transformer-based deep learning models for the sentiment analysis of social media data,” *Array*, vol. 14, p. 100157, Jul. 2022, doi: 10.1016/j.array.2022.100157.

[17] K. Kaushik and M. Parmar, “IMDb Movie Data Classification using Voting Classifier for Sentiment Analysis,” *Int. J. Comput. Sci. Eng.*, vol. 10, no. 1, pp. 18–23, Jan. 2022, doi: 10.26438/ijcse/v10i1.1823.

[18] P. Chakriswaran, D. R. Vincent, K. Srinivasan, V. Sharma, C.-Y. Chang, and D. G. Reina, “Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues,” *Appl. Sci.*, vol. 9, no. 24, p. 5462, Dec. 2019, doi: 10.3390/app9245462.

[19] Q. A. Xu, V. Chang, and C. Jayne, “A systematic review of social media-based sentiment analysis: Emerging trends and challenges,” *Decis. Anal. J.*, vol. 3, p. 100073, Jun. 2022, doi: 10.1016/j.dajour.2022.100073.

[20] S. Tripathi, R. Mehrotra, V. Bansal, and S. Upadhyay, “Analyzing Sentiment using IMDb Dataset,” in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhimtal, India: IEEE, Sep. 2020, pp. 30–33. doi: 10.1109/CICN49253.2020.9242570.

[21] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem, and T. Mahmood, “Sentiment analysis on IMDB using lexicon and neural networks,” *SN Appl. Sci.*, vol. 2, no. 2, p. 148, Feb. 2020, doi: 10.1007/s42452-019-1926-x.

[22] S. M. Y. Iqbal Tomal, “Sentiment Analysis of IMDb Movie Reviews,” *Int. J. Innov. Sci. Res. Technol. IJISRT*, pp. 2338–2343, Jun. 2024, doi: 10.38124/ijisrt/IJISRT24MAY1625.

[23] W. Ning, F. Wang, W. Wang, H. Wu, Q. Zhao, and T. Zhang, “Research on movie rating based on BERT-base model,” *Sci. Rep.*, vol. 15, no. 1, p. 9156, Mar. 2025, doi: 10.1038/s41598-025-92430-w.

[24] N. N. Marpid, Y. I. Kurniawan, and S. P. Rahayu, “ANALYSIS OF THE MOVIE DATABASE FILM RATING PREDICTION WITH ENSEMBLE LEARNING USING RANDOM FOREST REGRESSION METHOD,” *J. Tek. Inform. Jutif*, vol. 6, no. 1, pp. 1–10, Feb. 2025, doi: 10.52436/1.jutif.2025.6.1.1563.

[25] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, “The influence of preprocessing on text classification using a bag-of-words representation,” *PLOS ONE*, vol. 15, no. 5, p. e0232525, May 2020, doi: 10.1371/journal.pone.0232525.

[26] 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.

[27] D. Z. Abidin, S. Nurmaini, R. Firsandava Malik, Erwin, E. Rasywir, and Y. Pratama, “RSSI Data Preparation for Machine Learning,” in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia: IEEE, Nov. 2020, pp. 284–289. doi: 10.1109/ICIMCIS51567.2020.9354273.

[28] C. Petridis, “Text Classification: Neural Networks VS Machine Learning Models VS Pre-trained Models,” Dec. 30, 2024, *arXiv*: arXiv:2412.21022. doi: 10.48550/arXiv.2412.21022.

[29] R. V. K. Bevara, N. R. Mannuru, S. P. Karedla, and T. Xiao, “Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering,” *Appl. Sci.*, vol. 14, no. 8, p. 3483, Apr. 2024, doi: 10.3390/app14083483.

[30] M. Ni, Z. Sun, and W. Liu, “Reversible jump attack to textual classifiers with modification reduction,” *Mach. Learn.*, vol. 113, no. 9, pp. 5907–5937, Sep. 2024, doi: 10.1007/s10994-024-06539-6.

[31] M. Dörrich, M. Fan, and A. M. Kist, “Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks,” *IEEE Access*, vol. 11, pp. 57627–57634, 2023, doi: 10.1109/ACCESS.2023.3284388.

[32] V. Varadharajan, N. Smith, D. Kalla, F. Samaah, and V. Mandala, “Deep Learning-Based Sentiment Analysis: Enhancing IMDb Review Classification with LSTM Models,” *Univers. J. Comput. Sci. Commun.*, vol. 4, no. 1, pp. 1–14, Jan. 2025, doi: 10.31586/ujcsc.2025.1249.

[33] P. Atandoh, F. Zhang, M. A. Al-antari, D. Addo, and Y. Hyeon Gu, “Scalable deep learning framework for sentiment analysis prediction for online movie reviews,” *Heliyon*, vol. 10, no. 10, p. e30756, May 2024, doi: 10.1016/j.heliyon.2024.e30756.

[34] D. Z. Abidin, M. Rosario, and A. Sadikin, “Improving Term Deposit Customer Prediction Using Support Vector Machine with SMOTE and Hyperparameter Tuning in Bank Marketing Campaigns,” vol. 6, no. 3, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.3.4585.

[35] Z. J. Wang, R. Turko, and D. H. Chau, “Dodrio: Exploring Transformer Models with Interactive Visualization,” Jun. 05, 2021, *arXiv*: arXiv:2103.14625. doi: 10.48550/arXiv.2103.14625.

[36] A. M. P. Brasoveanu and R. Andonie, “Visualizing Transformers for NLP: A Brief Survey,” in *2020 24th International Conference Information Visualisation (IV)*, Melbourne, Australia: IEEE, Sep. 2020, pp. 270–279. doi: 10.1109/IV51561.2020.00051.

[37] F. Alzamzami and A. E. Saddik, “Transformer-Based Feature Fusion Approach for Multimodal Visual Sentiment Recognition Using Tweets in the Wild,” *IEEE Access*, vol. 11, pp. 47070–47079, 2023, doi: 10.1109/ACCESS.2023.3274744.