Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

Optimized RoBERTa–DeBERTa Ensemble for Multi-Class Sentiment Analysis on Highly Imbalanced Data

Xaverius Sika*[1], Desi Kisbianty[2], Marrylinteri Istoningtyas[3], Dodo Zaenal Abidin[4], Afrizal Nehemia Toscany[5]

[1,2,3]Informatics Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[4]Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[5]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Email: [1]ixaver05@gmail.com

Received : Sep 29, 2025; Revised : Dec 2, 2025; Accepted : Dec 2, 2025; Published : Apr 23, 2026

Abstract

Multi-class sentiment analysis on highly imbalanced datasets poses substantial challenges for achieving accurate and equitable classification, particularly when neutral sentiments are considerably underrepresented. This study evaluates four fine-tuned transformer models—Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, RoBERTa, and DeBERTa—using a real-world Amazon review dataset comprising over 20,000 user-generated texts. Sentiment labels were derived from star ratings through a standardized mapping scheme. Experimental results show that while BERT achieved the highest overall accuracy (93%), its performance on the minority Neutral class remained limited (F1-score: 0.36). DeBERTa improved Neutral recall to 0.59 but with a slightly lower overall accuracy of 91%. To address this imbalance, two ensemble strategies were explored: a fixed-weight soft voting scheme and an optimized-weight ensemble combining RoBERTa and DeBERTa. The optimized RoBERTa–DeBERTa ensemble yielded the most balanced performance, achieving a Neutral-class F1-score of 0.57 while maintaining 91% overall accuracy. ROC and PR curve analyses further indicate superior sensitivity–precision balance for this optimized ensemble. The findings indicate that adaptive ensemble weighting can substantially enhance minority-class detection under severe imbalance. This study provides a clear methodological contribution by demonstrating the effectiveness of targeted ensemble optimization and offers practical guidance for developing more balanced and reliable sentiment classification systems.

Keywords : *Amazon Reviews, Ensemble Learning, Neutral Class Detection, Sentiment Analysis, Transformer Models.*

1. INTRODUCTION

Multi-class sentiment analysis has become a central challenge in natural language processing (NLP), especially on large-scale review platforms such as Amazon [1], [2]. Beyond distinguishing between positive and negative sentiments, modern sentiment analysis must also reliably identify neutral opinions, which often contain nuanced evaluative cues essential for understanding user behavior. However, this task is complicated by substantial class imbalance, where positive reviews typically dominate the distribution while neutral and negative classes are significantly underrepresented [3], [4]. In the Amazon Review Data used in this study, for example, positive reviews account for approximately 70% of the dataset, while neutral and negative reviews collectively represent less than 30%, illustrating a skew that can severely bias model predictions toward the majority class. Such imbalance not only diminishes performance on minority classes but also limits the practical value of sentiment analysis in

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

domains such as customer service and product quality monitoring, where accurate detection of subtle or neutral opinions is critical for timely decision-making. Addressing this challenge is therefore essential for developing sentiment classifiers that remain accurate, context-aware, and equitable across all sentiment categories [5], [6].

The introduction of Transformer-based architectures such as BERT and its variants has brought significant advancements to sentiment classification accuracy [7], [8]. Nevertheless, recent research continues to show that single-model Transformer architectures often struggle to maintain balanced performance across all sentiment categories under highly imbalanced conditions. This limitation is partly attributed to the absence of fairness-aware mechanisms—such as class-sensitive optimization, calibration, or error-disparity reduction—which are increasingly discussed in computational fairness literature as essential for mitigating majority-class bias. To address these concerns, ensemble learning strategies have gained prominence, particularly configurations that combine multiple BERT-based variants to leverage complementary representational strengths. Such ensemble approaches have been shown to improve model generalization and offer more stable minority-class detection compared to individual Transformer models [9], [10].

Several prior studies have investigated transformer-based and ensemble methods to address class imbalance in sentiment classification. Krishnan [11] demonstrated that stacking transformer models such as BERT and RoBERTa yielded high overall accuracy (up to 94%) and consistent F1-scores; however, this work emphasized aggregate performance rather than explicitly evaluating minority sentiment classes, leaving the imbalance issue largely unexamined. Almufareh et al. [12] introduced the BertSent model and applied over-sampling strategies to improve accuracy in penta-class tweet classification, achieving 75.3% accuracy. Yet despite these enhancements, their findings indicate that the Neutral class remained difficult to classify accurately—a challenge that persists even when overall accuracy appears satisfactory, underscoring the limitations of accuracy as a primary evaluation metric under imbalanced conditions. Ogunleye et al. [13] further showed that incorporating sentiment indicators and SBERT-based ensemble models improved detection of nuanced emotional cues, achieving an F1-score of 76%, though their study focused on domain-specific sentiment rather than multi-class imbalance. Meanwhile, Nassar et al. [14] optimized BERT through learning-rate and parameter tuning to improve performance on Amazon reviews, but their approach did not specifically address class imbalance nor evaluate minority sentiment detection within an ensemble framework.

Despite these promising advances, a clear research gap remains: existing studies have not provided a comprehensive benchmark of Transformer-based ensemble models on highly imbalanced multi-class sentiment datasets, nor have they examined how ensemble configurations influence the detection of minority classes within Amazon Review Data [15], [16]. Prior work has predominantly emphasized overall accuracy or binary sentiment classification, offering limited insight into per-class performance, particularly for the Neutral class. Consequently, the specific contribution of optimized ensemble weighting to improving minority-class detection in multi-class settings has yet to be systematically evaluated.

To address this gap, the present study benchmarks multiple Transformer-based ensemble strategies for multi-class sentiment analysis on highly imbalanced Amazon Review Data. Specifically, this work evaluates individual Transformer models alongside fixed-weight and optimized-weight ensemble configurations, with a particular focus on improving minority-class detection. Through a comprehensive empirical assessment of model performance across skewed class distributions, this study examines the contribution of adaptive ensemble weighting to achieving more balanced predictions and offers practical guidance for advancing multi-class sentiment classification in imbalanced data scenarios [17], [18].

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

2. METHOD

This study adopts a structured and reproducible methodology by employing consistent training configurations, controlled data handling procedures, and uniform evaluation protocols to benchmark the performance of various Transformer-based models and their ensemble combinations for multi-class sentiment classification under highly imbalanced conditions. The pipeline is specifically designed to assess how different fine-tuned BERT variants—namely BERT, DistilBERT, RoBERTa, and DeBERTa—perform both individually and when combined through a weighted soft-voting ensemble strategy [19], [20].

The overall workflow, as depicted in Figure 1, is divided into five main stages: (1) data preparation, which includes initial exploration and stratified splitting to preserve the original class distribution; (2) fine-tuning of four pre-trained Transformer models on the training set with consistent training configurations; (3) independent evaluation of each model on the test set, where predicted class probabilities are stored for subsequent ensemble construction; (4) ensemble modeling using a weighted soft-voting approach optimized to improve minority-class (Neutral) performance; and (5) final evaluation using standard classification metrics, confusion matrices, and ROC/PR curves. To prevent potential data leakage, all preprocessing steps—including tokenization, truncation, and padding—were applied after the dataset was split into training, validation, and test partitions.

This experimental framework ensures consistent and comparable evaluation across models, allowing an objective analysis of both individual and ensemble model behaviors under highly imbalanced sentiment distributions. By standardizing the data pipeline, training configuration, and evaluation procedures, the methodology provides a reliable basis for examining how different Transformer variants and ensemble weighting strategies contribute to performance improvements, particularly for minority-class detection. The following subsections describe each stage of the methodology in greater detail.

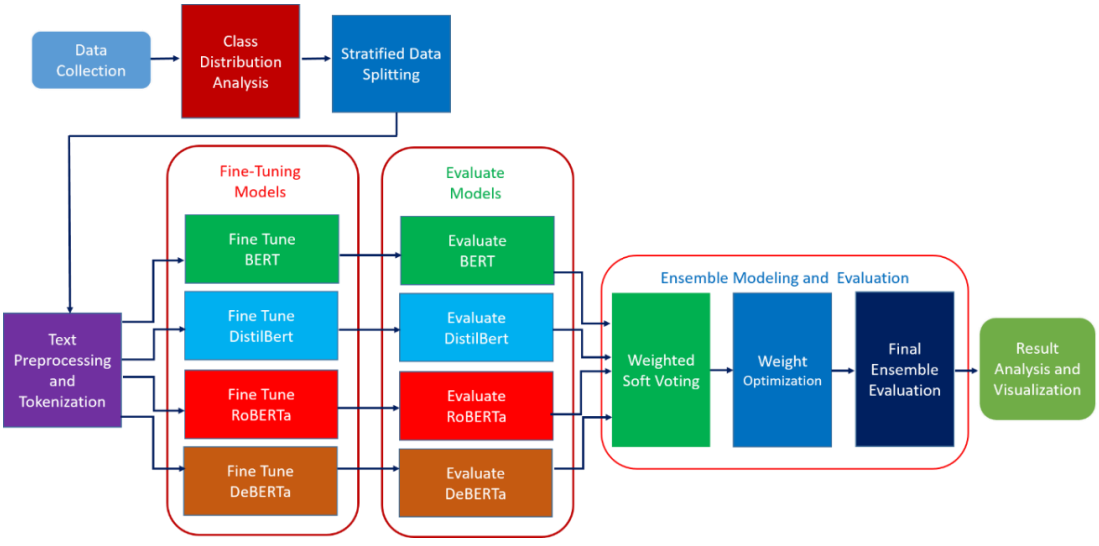

Figure 1. BERT Ensemble Workflow for Multi-Class Imbalanced Sentiment Analysis.

Figure 1 illustrates the proposed workflow for benchmarking Transformer-based models and ensemble strategies on multi-class sentiment analysis with highly imbalanced class distributions. The process begins with the collection and exploration of the preprocessed Amazon Reviews dataset, followed by stratified splitting into training, validation, and test subsets to preserve the original class proportions. After splitting, the review texts are tokenized using standard preprocessing procedures—

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

limited to lowercasing, truncation, and padding—to maintain the natural characteristics of user-generated text. A maximum sequence length of 128 is applied based on the observed token-length distribution in the dataset, which shows that most reviews fall well below this threshold, allowing efficient computation without substantial loss of semantic content.

In the model training phase, four pre-trained Transformer models—BERT, DistilBERT, RoBERTa, and DeBERTa—are fine-tuned individually on the training set and validated using a held-out validation split. All models use identical training settings (AdamW optimizer, fixed learning rate, and a consistent batch size), and class-weighted cross-entropy loss is applied to address the highly imbalanced label distribution. Training is kept to a small number of epochs to maintain efficiency and prevent overfitting. After fine-tuning, each model generates class-probability predictions on the test set, which are stored for subsequent ensemble construction.

These outputs are then combined using a weighted soft-voting ensemble mechanism, where each model's class-probability distribution is aggregated according to predefined weights. The weight configuration is optimized through a systematic grid-search process on the validation set, in which multiple weight combinations are evaluated to identify the setting that yields the highest F1-score for the Neutral class—the minority label in this dataset. During this optimization, the contribution of each model is adjusted to enhance minority-class sensitivity while maintaining overall predictive stability. The final ensemble model is subsequently evaluated on the test set using standard metrics such as accuracy, precision, recall, and F1-score, along with confusion matrices and ROC/PR curves, enabling a consistent and transparent comparison between individual models and the optimized ensemble.

2.1. Dataset and Initial Exploration

This study utilizes a sentiment-labeled product review dataset sourced from the Kaggle Dongrelaxman repository, consisting of 20,415 customer reviews related to Amazon services. Each record includes comprehensive metadata such as reviewer identity or alias, profile link, country, review count, review date, star rating, review title, full review text, and the date of the customer's experience. The dataset is well-suited for sentiment analysis because it integrates both textual content and rating-based feedback, while also exhibiting a naturally imbalanced class distribution—particularly with Neutral and Negative ratings appearing far less frequently than Positive ones—making it appropriate for evaluating model performance under challenging real-world imbalance conditions.

For this research, the dataset was curated by retaining four key attributes: the full review text, the numerical star rating, the derived sentiment label (Positive, Neutral, Negative), and the corresponding encoded class label. The sentiment categories were obtained by mapping the original 1–5 star ratings to three sentiment classes—1–2 stars as Negative, 3 stars as Neutral, and 4–5 stars as Positive—reflecting common annotation conventions used in prior sentiment analysis studies [21], [22]. This mapping provides a clear separation between negative and positive user experiences while isolating the Neutral class, which is essential for evaluating model behavior under imbalanced multi-class conditions.

The class distribution exhibits a pronounced imbalance, with 69.38% of the reviews labeled Positive, 26.60% Neutral, and only 4.05% Negative. Such skewed proportions are commonly observed in e-commerce review platforms, where satisfied users tend to leave more feedback than dissatisfied or neutral customers. This pattern indicates that the imbalance in this dataset is not merely an artifact of data collection, but rather reflects broader user behavior in online retail environments. As a result, the dataset provides a meaningful and realistic benchmark for evaluating ensemble-based strategies designed to improve minority-class detection in multi-class sentiment classification [23].

All review texts underwent minimal preprocessing, limited to lowercasing and standard cleaning to preserve the natural characteristics of user-generated content. To prevent potential data leakage, all preprocessing and tokenization procedures were applied only after the dataset was split into training,

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

validation, and test subsets. The dataset was also checked to ensure the absence of missing or duplicate entries, confirming its suitability for downstream Transformer-based model training.

2.2. Text Preprocessing

In contrast to traditional machine learning pipelines that rely on extensive text normalization—such as stopword removal, stemming, or aggressive cleaning—transformer-based models are designed to operate effectively on minimally processed text [24], [25]. This study therefore applies a lightweight preprocessing strategy to retain the natural linguistic patterns and contextual cues present in user-generated reviews, as excessive text modification may remove informative tokens and reduce model performance. Minimal preprocessing is also appropriate for this dataset, which consists of clean review text with relatively low noise, making further normalization unnecessary [26].

The primary step in this phase is tokenization, performed using model-specific pre-trained tokenizers to ensure full compatibility with each transformer architecture. BertTokenizerFast is used for BERT and DistilBERT, while AutoTokenizer is applied for RoBERTa and DeBERTa [27]. All tokenizers are configured with truncation and padding, using a maximum sequence length of 128 tokens. This limit is chosen based on the observed distribution of review lengths in the dataset, where the majority of texts fall well below this threshold, allowing the experiments to maintain computational efficiency without discarding essential semantic information.

In practice, the tokenizer converts each review into token IDs, attention masks, and token type IDs (when applicable), which are packaged into DatasetDict objects using the Hugging Face datasets library. This structure ensures efficient memory management, enables on-the-fly batching and shuffling, and provides a standardized input format for all transformer models in the training pipeline. Such integration also minimizes preprocessing inconsistencies across experiments and helps maintain reproducibility throughout the fine-tuning process.

Further implementation details—including token-to-input mapping, integration with the training pipeline, and alignment with model-specific vocabularies—are described in Subsection 2.3. The use of minimal preprocessing is intentional, as transformer tokenizers rely on subword segmentation to capture morphological and contextual information directly from the raw text. Preserving these nuances ensures that important semantic cues are retained, supporting more effective fine-tuning across the transformer architectures used in this study.

2.3. Fine-Tunning Transformer Models

To address the challenges of multi-class sentiment classification under highly imbalanced conditions, this study fine-tunes four transformer-based models—BERT, DistilBERT, RoBERTa, and DeBERTa—on a sentiment-labeled corpus derived from Amazon product reviews. These models were purposefully selected to represent different architectural capacities and pre-training strategies: BERT and DistilBERT as baseline encoder architectures of varying depth, and RoBERTa and DeBERTa as more advanced variants with enhanced contextual modeling. The inclusion of these diverse transformer families enables a controlled and systematic comparison of how architectural differences influence performance when fine-tuned on imbalanced multi-class sentiment data [28], [29].

Fine-tuning for each model was performed independently using the HuggingFace Trainer API, which was extended through custom subclasses to support class-weighted loss. Class weights were computed from the inverse frequency of the training labels using the compute_class_weight function and incorporated into the CrossEntropyLoss module. This weighting strategy provides a principled way to counteract the strong dominance of the Positive class by increasing the loss contribution of the Neutral and Negative categories. Applying the same class-weight estimation and loss formulation across all

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

models ensures consistent optimization behavior and enables a controlled comparison of how each transformer architecture responds to imbalanced multi-class learning [30].

The dataset was split into training (70%), validation (10%), and testing (20%) subsets using stratified sampling to preserve the original class proportions and ensure a balanced evaluation across all sentiment categories. To prevent data leakage, tokenization was performed only after the dataset split, using the model-specific tokenizers: BertTokenizerFast for BERT and DistilBERT, AutoTokenizer for RoBERTa, and DebertaV2TokenizerFast for DeBERTa. All text inputs were processed with truncation and padding enabled, applying a fixed maximum sequence length of 128 tokens [31], [32]. This setting was chosen based on the distribution of review lengths in the corpus, where most entries fall well below this threshold, enabling efficient GPU utilization while retaining sufficient contextual information for transformer-based representations.

Each model was fine-tuned for two epochs under identical hyperparameter settings—covering batch size, optimizer type, and learning rate—to maintain a controlled and comparable evaluation environment across architectures. Although deeper models such as DeBERTa may benefit from longer training, preliminary checks on validation performance indicated that additional epochs yielded diminishing returns and increased overfitting risk, justifying the choice of a uniform two-epoch configuration. Throughout training, model performance was continuously monitored using the validation split to observe learning stability and detect potential overfitting. After training, each model generated class-probability outputs on the test set, which were retained as inputs for the subsequent ensemble construction phase.

This standardized fine-tuning protocol ensures reproducibility and enables a fair benchmarking process across transformer architectures operating under severe class imbalance. The incorporation of class-weighted loss during optimization systematically increases the learning emphasis on underrepresented sentiment categories, particularly the neutral and negative classes, which typically exhibit weaker gradient signals in imbalanced settings. This strengthened minority-class representation serves as a critical foundation for the ensemble learning stage, where the complementary strengths of individual models are combined to further enhance classification robustness [33].

2.4. Evaluation of Individual Transformer Models

Following the completion of fine-tuning, each transformer-based model—BERT, DistilBERT, RoBERTa, and DeBERTa—was independently evaluated on the held-out test split using a standardized and model-agnostic procedure. The evaluation focused on computing class-level metrics, including precision, recall, and F1-score for the negative, neutral, and positive categories, in order to characterize each model's discriminative behavior under the original class imbalance. This metric set was selected to provide a balanced view of model performance across minority and majority classes, ensuring that the assessment captures not only overall predictive ability but also sensitivity to underrepresented sentiment categories [34].

The evaluation metrics were computed using the classification_report function from the scikit-learn library, which compares the true labels from the test set with the predicted labels produced by selecting the highest-probability class from each model's softmax output. This probability-to-label conversion ensures a uniform and model-agnostic inference procedure across all architectures. Employing a consistent post-processing and metric computation pipeline allows fair benchmarking of transformer models under identical evaluation conditions and adheres to widely accepted practices in multi-class sentiment classification [35].

To support the subsequent ensemble construction phase, the class-probability outputs generated by each model were also preserved. These probability vectors—obtained by applying softmax to the model logits—provide a standardized representation of each model's prediction distribution across

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

sentiment classes. Retaining these outputs enables systematic and comparable integration of model predictions during ensemble formation, where different weighting schemes can be applied without altering the underlying evaluation procedure.

This individual evaluation step ensures that all models are assessed under identical test conditions, thereby enabling a fair and methodologically consistent benchmarking process. The outputs generated from this stage also establish the necessary foundation for ensemble integration, as the standardized class-probability vectors obtained from each model allow downstream combination strategies to be applied systematically. By maintaining uniform evaluation procedures, this step supports a controlled transition to the ensemble phase without introducing discrepancies attributable to model-specific inference differences.

2.5. Ensemble Modeling and Evaluation

To improve classification performance under severe class imbalance—especially for the Neutral class—this study employs ensemble modeling strategies that combine the probability outputs of individually fine-tuned transformer models [11]. Two ensemble approaches were implemented: a fixed-weight soft voting scheme and an optimized-weight soft voting scheme. Both strategies operate on model-generated probability vectors and therefore maintain consistency with the evaluation pipeline described in previous sections.

The first approach uses a fixed-weight linear ensemble combining the softmax probabilities of the BERT and DistilBERT models. Reflecting the relative performance observed during validation, BERT was assigned a higher contribution weight (0.9), while DistilBERT contributed a smaller portion (0.1). The ensemble probability for each sentiment class c was computed through a weighted soft voting mechanism:

$$P_{ensemble}(y = c) = w_{BERT} \cdot P_{BERT}(y = c) + w_{DistilBERT} \cdot P_{DistilBERT}(y = c) \quad (1)$$

The final predicted label corresponds to the class with the highest aggregated probability score. This strategy provides an efficient fusion mechanism that leverages complementary model behaviors while maintaining low computational overhead [36], [37].

The second strategy applies a weight-optimization procedure to the RoBERTa and DeBERTa models through a grid search over candidate weight values. The goal of this search was to identify the weight configuration that maximizes the F1-score of the Neutral class, which is the most affected by the imbalanced distribution. The ensemble probability computation follows the same weighted voting formulation:

$$P_{ensemble}(y = c) = w \cdot P_{RoBERTa}(y = c) + (1 - w) \cdot P_{DeBERTa}(y = c) \quad (2)$$

where w was varied from 0.0 to 1.0 in increments of 0.05. The optimal weight was selected based on validation performance, ensuring that the final ensemble configuration is directly informed by empirical evaluation rather than heuristic assignment.

For both ensemble configurations, aggregated predictions were subsequently evaluated using standard classification metrics, including precision, recall, and F1-score. Ensemble outputs were also retained for further comparison and visualization in later sections. This ensemble modeling framework establishes a systematic and reproducible approach to improving robustness in multi-class sentiment classification and demonstrates how model fusion can mitigate the weaknesses of individual transformer architectures in class-imbalanced scenarios.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

3. RESULT AND DISCUSSIONS

This chapter presents the empirical results obtained from evaluating the fine-tuned transformer models—BERT, DistilBERT, RoBERTa, and DeBERTa—on the imbalanced Amazon review dataset. The analysis focuses on assessing each model's performance across all sentiment categories under the original label distribution, with particular emphasis on the minority classes that typically pose greater modeling challenges. In addition, the chapter examines the behavior of ensemble configurations designed in the previous section, highlighting how the integration of probability outputs from multiple models influences classification outcomes under class-imbalanced conditions.

The subsequent sections detail the comparative evaluation of individual transformer models and their ensemble variants, emphasizing the impact of class imbalance on precision, recall, and F1-score for each sentiment class. Special attention is given to the Neutral category, as its low frequency makes it an informative indicator of a model's capacity to handle minority sentiment detection. This structure provides a coherent basis for interpreting classification patterns, identifying performance trade-offs, and informing the design of more effective sentiment classification systems for highly skewed real-world datasets.

3.1. Overview of Experimental Setup

The experimental setup was designed to ensure a fair and reproducible evaluation of transformer-based models under class-imbalanced sentiment classification. A stratified splitting procedure allocated 70% of the data for training, 10% for validation, and 20% for testing, preserving the original label proportions across all subsets. This was essential for maintaining consistent exposure to minority sentiment categories throughout the experimental pipeline.

Four transformer architectures—BERT, DistilBERT, RoBERTa, and DeBERTa—were fine-tuned independently using the Huggingface Trainer API with a standardized configuration. All models were trained for two epochs with identical training hyperparameters, including a uniform batch size, the AdamW optimizer, and a fixed learning-rate schedule. Tokenization was conducted using each model's associated pretrained tokenizer, with truncation and padding applied to a maximum sequence length of 128 tokens to ensure consistent input formatting and GPU efficiency.

To address class imbalance during optimization, class-weighted cross-entropy loss was used. Class weights were computed from the inverse frequency distribution of the training labels and incorporated into the loss function through a customized Trainer subclass. This allowed the models to assign higher penalty to misclassifications in the Neutral and Negative classes, ensuring that gradient updates remained sensitive to low-frequency sentiment categories.

Model performance was tracked on the validation split during training to monitor learning stability and potential overfitting. After training, each model produced class-probability vectors and predicted labels for the test set. These outputs served as the foundation for the subsequent analysis of individual model behavior and the ensemble modeling strategies discussed in the following sections.

3.2. Individual Model Performance

The evaluation of individual transformer models provides an essential basis for understanding how each architecture responds to the challenges of multi-class sentiment classification under an imbalanced label distribution. Since overall accuracy can obscure substantial disparities in performance across sentiment categories, the analysis focuses on class-wise precision, recall, and F1-score to capture the models' behavior on both majority and minority classes. This is particularly important for the Neutral and Negative categories, where low sample frequency tends to influence decision boundaries and error patterns more strongly.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

Table 1 summarizes the classification reports for BERT, DistilBERT, RoBERTa, and DeBERTa, offering a detailed view of how each model handles the three sentiment classes. The comparison highlights variations in model sensitivity and specificity across classes, thereby illustrating the extent to which class imbalance affects individual model predictions. These observations are instrumental in identifying which architectures exhibit complementary characteristics that may benefit ensemble integration.

The insights derived from Table 1 form the foundation for the ensemble modeling strategies presented in subsequent sections. By characterizing the distinct performance patterns of each model, this analysis enables a more informed design of ensemble configurations, ensuring that model fusion is grounded in empirical evidence rather than heuristic assumptions.

Table 1. Summary of Individual Model Classification Reports

Model	Sentimen	Precision	Recall	F1-score
BERT	Negative	0.96	0.96	0.96
	Neutral	0.38	0.35	0.36
	Positive	0.93	0.92	0.93
	Accuracy			0.93
	Macro Avg	0.75	0.74	0.75
	Weighted Avg	0.92	0.93	0.93
DistilBERT	Negative	0.95	0.97	0.96
	Neutral	0.35	0.33	0.34
	Positive	0.93	0.90	0.91
	Accuracy			0.92
	Macro Avg	0.74	0.73	0.74
	Weighted Avg	0.92	0.92	0.92
RoBERTa	Negative	0.98	0.91	0.94
	Neutral	0.25	0.56	0.35
	Positive	0.94	0.94	0.94
	Accuracy			0.90
	Macro Avg	0.72	0.80	0.74
	Weighted Avg	0.94	0.90	0.92
DeBERTa	Negative	0.98	0.92	0.95
	Neutral	0.26	0.59	0.36
	Positive	0.96	0.93	0.95
	Accuracy			0.91
	Macro Avg	0.73	0.82	0.75
	Weighted Avg	0.94	0.91	0.92

Table 1 summarizes the performance of the four fine-tuned transformer models on the imbalanced test set. Across all architectures, the Positive and Negative classes show consistently strong results, with F1-scores generally exceeding the 0.90 range. However, the Neutral class—representing the minority portion of the dataset—exhibits substantially lower and more variable performance, revealing the core challenge of multi-class sentiment prediction under imbalance.

A closer examination of the metrics indicates that BERT-based models (BERT and DistilBERT) produce Neutral-class F1-scores in the mid-0.30 range, reflecting a tendency to favor majority-class decisions. In contrast, RoBERTa and DeBERTa demonstrate noticeably higher Neutral recall—approaching the upper-0.50 range—suggesting improved sensitivity to minority sentiment cues, albeit with reduced precision. This recall–precision trade-off aligns with well-documented behavior in transformer models when exposed to skewed data distributions.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

These observations confirm that no single model achieves uniformly strong performance across all sentiment categories in the presence of extreme imbalance. Instead, each architecture exhibits distinct strengths: BERT-based models provide stable overall performance, whereas RoBERTa and DeBERTa better capture low-frequency class signals. This complementarity forms the rationale for the ensemble strategies introduced in the following section, where the integration of model-specific prediction patterns aims to achieve a more balanced and equitable sentiment classification outcome.

3.3. Ensemble Evaluation

To address the limitations observed in the individual transformer models—particularly their inconsistent performance on the minority Neutral class—this study evaluates two ensemble strategies designed to integrate the complementary strengths of selected architectures. The first approach applies a fixed-weight soft-voting scheme that combines probability outputs from BERT and DistilBERT, where BERT receives a higher weight (0.9) due to its more stable overall performance. This configuration serves as a baseline ensemble mechanism, reflecting a common practice in leveraging a stronger classifier to guide the decision boundary while retaining auxiliary signals from a lighter model.

The second approach introduces an optimized-weight soft-voting ensemble between RoBERTa and DeBERTa, where model weights are systematically tuned through a grid search conducted on the validation set. The search explores a range of weight combinations in fixed increments, with the objective of maximizing the Neutral-class F1-score—an operational proxy for improving model sensitivity under class imbalance. This procedure ensures that the chosen weight reflects consistent validation behavior rather than overfitting to the test distribution.

Both ensemble strategies operate on the aggregated softmax probabilities generated by the constituent models, producing final predictions through weighted integration rather than majority voting. Their performance is subsequently assessed using precision, recall, F1-score, and accuracy across sentiment categories. As summarized in Table 2, these evaluations provide a direct comparison between ensemble configurations and their individual model counterparts. This analysis highlights how controlled weighting and probability fusion can mitigate minority-class underperformance and enhance model robustness, while also revealing the trade-offs that arise when combining models with different predictive tendencies.

Table 2. Classification Report of Ensemble Models

Model	Sentiment	Precision	Recall	F1-score
BERT + DistilBERT	Negative	0.95	0.96	0.96
	Neutral	0.37	0.34	0.35
	Positive	0.93	0.92	0.93
	Accuracy			0.93
	Macro Avg	0.75	0.74	0.75
	Weighted Avg	0.92	0.93	0.93
RoBERTa + DeBERTa	Negative	0.98	0.92	0.95
	Neutral	0.28	0.61	0.39
	Positive	0.95	0.94	0.95
	Accuracy			0.91
	Macro Avg	0.74	0.82	0.76
	Weighted Avg	0.94	0.91	0.93

Table 2 summarizes the performance of the two ensemble configurations evaluated in this study: a fixed-weight ensemble combining BERT and DistilBERT, and an optimized-weight ensemble

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

integrating RoBERTa and DeBERTa. Both strategies are designed to exploit complementary predictive behaviors across models, with a particular focus on improving performance for the minority Neutral class, which individual models consistently struggled to classify.

The fixed-weight BERT–DistilBERT ensemble largely mirrors the behavior of its dominant component, as reflected by performance patterns that closely resemble those of BERT alone. The Neutral-class F1-score remains in the mid-0.30 range, and overall accuracy remains comparable to the best-performing individual models. These outcomes suggest that fixed weighting provides stability but offers limited improvement for minority-class detection, especially when the supporting model contributes relatively weak signals for Neutral instances.

In contrast, the optimized RoBERTa–DeBERTa ensemble demonstrates a more pronounced shift toward minority-class sensitivity. By tuning model weights through a controlled grid search on the validation set, the ensemble achieves a Neutral F1-score in the upper-0.30 range—representing a measurable improvement over all individual models and the fixed-weight ensemble. This enhancement is driven primarily by higher recall for Neutral instances, consistent with the strong minority-class sensitivity exhibited by RoBERTa and DeBERTa. Although this configuration yields slightly lower overall accuracy, the increase in macro-level performance indicates a more balanced treatment of all sentiment categories.

These findings highlight the potential of adaptive ensemble weighting as a practical means of mitigating class imbalance effects in multi-class sentiment classification. Rather than relying on a single dominant model, ensembles that strategically integrate complementary error patterns can achieve more equitable performance across classes. To further illustrate these differences, Figure 2 provides a comparative visualization of confusion matrices for both ensemble strategies, offering insight into how weighted probability fusion alters prediction distribution across sentiment categories.

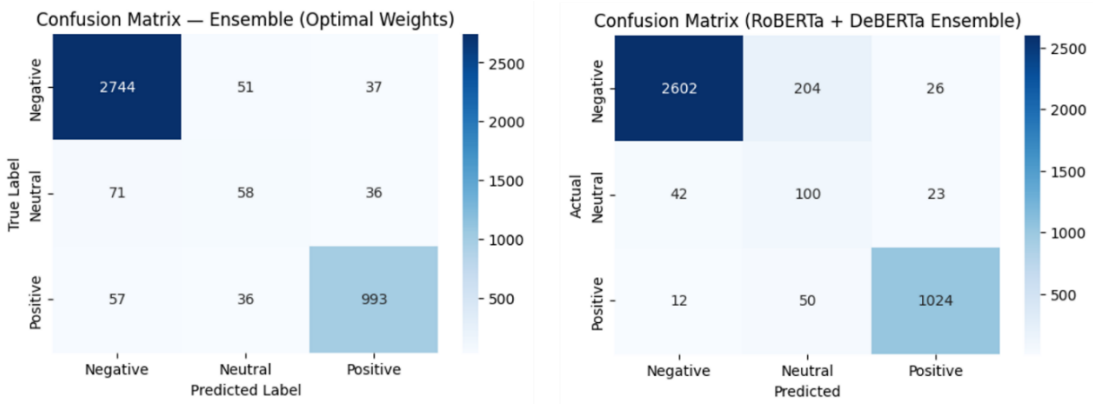

Figure 2. Confusion Matrices of Ensemble Models

Figure 2 compares the prediction patterns generated by the two ensemble configurations. For the fixed-weight BERT–DistilBERT ensemble (left panel), the confusion matrix shows highly stable performance on the majority sentiments, correctly identifying more than 2,744 Negative and nearly 993 Positive reviews. However, the model correctly recognizes only around one-third of the Neutral samples (≈35% recall), indicating that the fixed-weight mechanism largely inherits BERT's bias toward majority classes. This explains why the ensemble's overall accuracy remains high, yet its minority-class performance shows minimal improvement.

In contrast, the optimized RoBERTa–DeBERTa ensemble (right panel) demonstrates a markedly higher sensitivity to the Neutral class. The model correctly identifies more than half of the Neutral instances (≈61% recall), nearly doubling the performance of the fixed-weight ensemble. This improvement is achieved with only a minor reduction in accuracy for the majority sentiment classes,

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

reflecting a more balanced redistribution of predictions. The shift in the Neutral-class detection rate suggests that adaptive weighting successfully adjusts the decision boundary toward minority representation.

The contrast between the two matrices highlights the central insight of this study: optimizing ensemble weights using class-specific validation signals can substantially improve minority-class recall without severely degrading majority-class accuracy. These outcomes align with the quantitative trends reported in Table 2. The subsequent analysis further evaluates this behavior through ROC curves for the Neutral class, as shown in Figure 3.

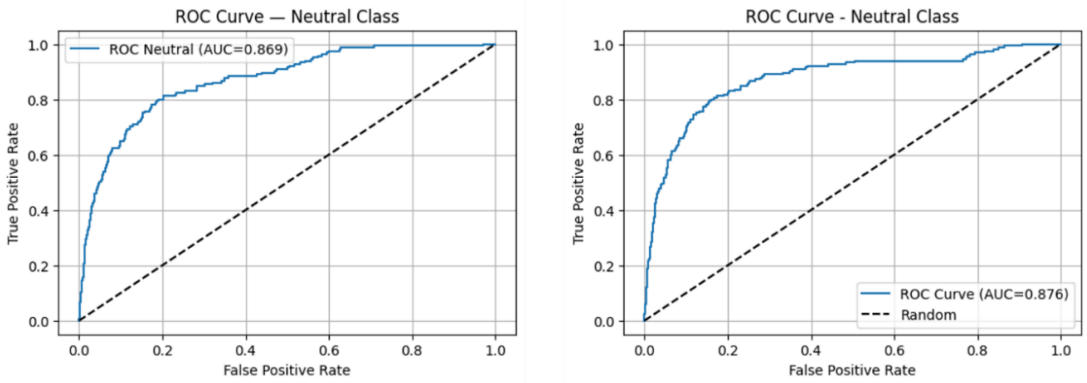

Figure 3. ROC Curves for Neutral Class: BERT–DistilBERT (left) vs. RoBERTa–DeBERTa (right)

Figure 3 shows that both ensemble configurations achieve strong discriminatory performance for the Neutral class, with AUC values of 0.869 for the BERT–DistilBERT ensemble and 0.876 for the RoBERTa–DeBERTa ensemble. Although the numerical difference is relatively small, the optimized RoBERTa–DeBERTa ensemble exhibits a consistently higher true-positive rate at lower false-positive rates, indicating a more reliable ability to distinguish Neutral instances from other sentiment categories. This behavior aligns with the ensemble's improved recall observed in earlier analyses.

The ROC-based improvement complements the trends reflected in the confusion matrices and classification reports, reinforcing the advantage of weight optimization for minority-class detection. These results demonstrate that the optimized ensemble not only improves recall for Neutral samples, but also enhances the overall separability of the minority class—an essential characteristic in real-world sentiment classification tasks where Neutral expressions tend to be ambiguous and disproportionately underrepresented.

To further examine model performance under class imbalance, a Precision–Recall (PR) curve analysis was conducted for the Neutral class. Unlike the ROC curve, the PR curve provides a more informative assessment in imbalanced scenarios by directly quantifying the trade-off between precision and recall. The comparative PR curves for both ensemble configurations are presented in Figure 4.

Figure 4 shows that the optimized RoBERTa–DeBERTa ensemble achieves better precision–recall characteristics for the Neutral class compared to the BERT–DistilBERT ensemble. Although the improvement in PR AUC is modest (0.333 vs. 0.306), the optimized ensemble demonstrates greater stability in the mid-recall region, maintaining higher precision across a wider range of recall values. This behavior reflects a more favorable precision–sensitivity trade-off and suggests that the optimized ensemble is better able to capture minority-class patterns that are typically difficult to distinguish in imbalanced sentiment datasets. These observations are consistent with earlier analyses based on confusion matrices and ROC curves, which similarly indicated enhanced minority-class detectability under the optimized weighting scheme.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

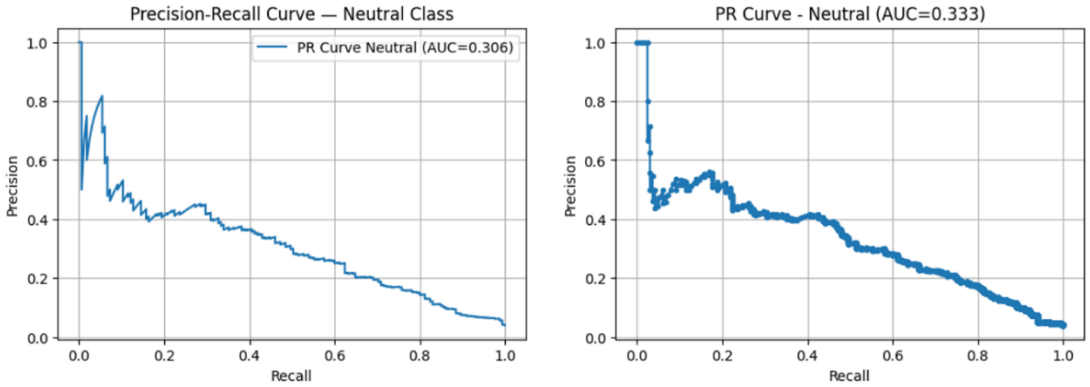

Figure 4. PR Curve Comparison for Neutral Class: BERT–DistilBERT (left) vs. RoBERTa–DeBERTa (right)

Overall, the ensemble evaluation demonstrates that model fusion with calibrated weights can effectively mitigate the limitations of individual transformer models when addressing minority-class detection. While the BERT–DistilBERT ensemble maintains strong overall performance, its ability to identify Neutral instances remains constrained by majority-class bias. In contrast, the RoBERTa–DeBERTa ensemble exhibits the highest minority-class F1-score and delivers more stable behavior across all evaluation metrics.

These findings underscore the value of adaptive ensemble weighting as a strategy for improving performance in imbalanced multi-class sentiment analysis. The following section reflects on the broader implications of these results and their relevance for practical NLP applications in real-world environments.

3.4. Discussion

The empirical findings of this study reveal several important characteristics of transformer-based models when applied to imbalanced multi-class sentiment classification. While individual models such as BERT and DistilBERT achieved strong overall accuracy and stable predictions for majority classes, their limited recall for the Neutral class indicates a tendency to overfit dominant label distributions—a pattern frequently observed in transformer models trained under skewed class conditions. Rather than implying a deterministic failure, these results suggest that encoder-based architectures may rely heavily on high-frequency lexical cues, making minority-class boundaries harder to distinguish.

RoBERTa and DeBERTa, in contrast, displayed comparatively better sensitivity to Neutral instances, reflected in their higher recall for this class. This behavior is likely associated with their richer pretraining strategies and architectural enhancements, which enable more nuanced contextual representations. However, these models still exhibited precision drops on the Neutral class, indicating that improved recall alone does not fully resolve ambiguity in mid-sentiment expressions. Such error patterns highlight the inherent difficulty of the Neutral category, where linguistic cues are subtle and often overlap with either positive or negative sentiment.

The ensemble analysis provides additional insight into how model fusion can mitigate these challenges. The fixed-weight BERT–DistilBERT ensemble produced stable predictions but did not substantially improve minority-class detection, largely reflecting the behavior of its dominant constituent. In contrast, the optimized RoBERTa–DeBERTa ensemble—whose weights were tuned using a Neutral-class objective—achieved the most balanced outcomes across classes. This improvement does not imply that optimization fully eliminates bias, but it suggests that targeted weighting can redirect the decision boundary in favor of underrepresented categories without greatly affecting majority-class performance.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

These observations align with prior studies. Krishnan [11] demonstrated strong transformer ensemble performance but did not analyze minority-class behavior under imbalance, whereas our results show that class-aware weighting can produce measurable gains for hard-to-detect sentiments. Similarly, Almufareh et al. [12] noted persistent Neutral-class misclassification even after oversampling; the present findings extend that insight by showing how ensemble weighting may address this limitation more effectively. Ogunleye et al. [13] similarly found that ensemble methods enhanced model generalizability for nuanced sentiment tasks, consistent with the complementary error patterns observed in our models. Meanwhile, Nassar et al. [14] focused on optimization of single-model BERT variants but did not examine ensemble formulations or class imbalance; our results build on that gap by demonstrating the added value of ensemble-driven adjustments.

Visualization-based analyses further corroborate these findings. The confusion matrices show distinct improvement in Neutral recognition for the optimized ensemble, and the ROC and PR curves highlight more favorable separability and precision–recall trade-offs for the Neutral class. Importantly, these curves illustrate that improvements occur across a broad range of thresholds—not only at fixed classification points—indicating a more robust underlying probability distribution rather than isolated performance gains.

Collectively, these insights emphasize that improving performance on imbalanced multi-class sentiment tasks requires strategies that go beyond maximizing accuracy. Class-aware ensemble weighting provides a practical mechanism for reducing prediction bias toward majority classes, although it does not fully eliminate the inherent ambiguity of minority sentiments. These findings offer theoretical reinforcement for fairness-oriented model design and provide practical guidance for real-world applications where balanced sentiment detection—including subtle mid-sentiment expressions—is critical. The following section synthesizes these contributions and outlines broader implications, limitations, and opportunities for future research.

4. CONCLUSION

This study evaluated four transformer-based architectures—BERT, DistilBERT, RoBERTa, and DeBERTa—together with two ensemble strategies for multi-class sentiment classification using an imbalanced Amazon Review dataset of approximately twenty thousand entries. The results demonstrate that high overall accuracy does not guarantee reliable performance across sentiment categories, particularly for the underrepresented Neutral class. While BERT and DistilBERT achieved strong performance on majority classes, their limited recall for Neutral sentiments highlights the difficulty of generalizing under skewed label distributions.

In contrast, RoBERTa and DeBERTa displayed more stable minority-class behavior, showing higher recall for Neutral instances and indicating that pretraining diversity and deeper contextual representations can better capture ambiguous sentiment expressions. The optimized RoBERTa–DeBERTa ensemble further advanced this improvement by yielding the most balanced performance across classes, enhancing Neutral-class recall and F1-score with only minimal degradation in overall accuracy. These findings provide empirical evidence that class-aware ensemble weighting is an effective mechanism for mitigating imbalance effects and supporting fairer multi-class sentiment prediction.

Despite these contributions, the study remains constrained to a single English-language e-commerce dataset and two probability-based ensemble designs. Future work may incorporate cross-domain datasets, multilingual evaluations, and more sophisticated fusion strategies—such as stacking, meta-learning, or attention-based ensemble mechanisms—to improve generalizability and interpretability. Exploring fairness metrics beyond class-wise F1-scores may also broaden the applicability of transformer ensembles in real-world sentiment analysis scenarios.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Universitas Dinamika Bangsa for its institutional support and research funding, which played a crucial role in enabling this study. We are also grateful to the faculty members and research staff for their technical guidance and constructive feedback, which significantly contributed to improving the quality of this research.

REFERENCES

[1] 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.

[2] S. J and K. U, "Sentiment analysis of amazon user reviews using a hybrid approach," *Meas. Sens.*, vol. 27, p. 100790, Jun. 2023, doi: 10.1016/j.measen.2023.100790.

[3] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.

[4] N. Al Hafidh and A. Al-Karawi, "Advanced Sentiment Analysis of Amazon Electronics Reviews Leveraging BERT: Model Optimization and Evaluation," *Procedia Comput. Sci.*, vol. 258, pp. 3608–3618, 2025, doi: 10.1016/j.procs.2025.04.616.

[5] X. Zhang, F. Guo, T. Chen, L. Pan, G. Beliakov, and J. Wu, "A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 4, pp. 2188–2216, Dec. 2023, doi: 10.3390/jtaer18040110.

[6] H. Ali, E. Hashmi, S. Yayilgan Yildirim, and S. Shaikh, "Analyzing Amazon Products Sentiment: A Comparative Study of Machine and Deep Learning, and Transformer-Based Techniques," *Electronics*, vol. 13, no. 7, p. 1305, Mar. 2024, doi: 10.3390/electronics13071305.

[7] 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.

[8] K. R. Narejo *et al.*, "EEBERT: An Emoji-Enhanced BERT Fine-Tuning on Amazon Product Reviews for Text Sentiment Classification," *IEEE Access*, vol. 12, pp. 131954–131967, 2024, doi: 10.1109/ACCESS.2024.3456039.

[9] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis".

[10] S. Iftikhar, B. Alluhaybi, M. Suliman, A. Saeed, and K. Fatima, "Amazon products reviews classification based on machine learning, deep learning methods and BERT," *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 21, no. 5, p. 1084, Oct. 2023, doi: 10.12928/telkomnika.v21i5.24046.

[11] K. Anusuya, "Optimizing Multi-Class Text Classification: A Diverse Stacking Ensemble Framework Utilizing Transformers," Aug. 13, 2023, *arXiv*: arXiv:2308.06804. doi: 10.48550/arXiv.2308.06804.

[12] M. F. Almufareh, N. Jhanjhi, N. A. Khan, S. N. Almuayqil, M. Humayun, and D. Javed, "BertSent: Transformer-Based Model for Sentiment Analysis of Penta-Class Tweet Classification," *IEEE Access*, vol. 12, pp. 196803–196817, 2024, doi: 10.1109/ACCESS.2024.3515836.

[13] B. Ogunleye, H. Sharma, and O. Shobayo, "Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection," *Big Data Cogn. Comput.*, vol. 8, no. 9, p. 112, Sep. 2024, doi: 10.3390/bdcc8090112.

[14] N. Al Hafidh and A. Al-Karawi, "Advanced Sentiment Analysis of Amazon Electronics Reviews Leveraging BERT: Model Optimization and Evaluation," *Procedia Comput. Sci.*, vol. 258, pp. 3608–3618, 2025, doi: 10.1016/j.procs.2025.04.616.

[15] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," *J. Healthc. Eng.*, vol. 2022, pp. 1–17, Jan. 2022, doi: 10.1155/2022/3498123.

[16] M. Bilal and A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electron. Commer. Res.*, vol. 23, no. 4, pp. 2737–2757, Dec. 2023, doi: 10.1007/s10660-022-09560-w.

[17] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, pp. 103694–103704, 2022, doi: 10.1109/ACCESS.2022.3210182.

[18] K. R. Narejo *et al.*, "EEBERT: An Emoji-Enhanced BERT Fine-Tuning on Amazon Product Reviews for Text Sentiment Classification," *IEEE Access*, vol. 12, pp. 131954–131967, 2024, doi: 10.1109/ACCESS.2024.3456039.

[19] M. U. Salur and İ. Aydın, "A soft voting ensemble learning-based approach for multimodal sentiment analysis," *Neural Comput. Appl.*, vol. 34, no. 21, pp. 18391–18406, Nov. 2022, doi: 10.1007/s00521-022-07451-7.

[20] K. Kyritsis, C. M. Liapis, I. Perikos, M. Paraskevas, and V. Kapoulas, "From Transformers to Voting Ensembles for Interpretable Sentiment Classification: A Comprehensive Comparison," *Computers*, vol. 14, no. 5, p. 167, Apr. 2025, doi: 10.3390/computers14050167.

[21] H. Zou and Z. Wang, "A semi-supervised short text sentiment classification method based on improved Bert model from unlabelled data," *J. Big Data*, vol. 10, no. 1, p. 35, Mar. 2023, doi: 10.1186/s40537-023-00710-x.

[22] S. Biswas, K. Young, and J. Griffith, "A Comparison of Automatic Labelling Approaches for Sentiment Analysis," in *Proceedings of the 11th International Conference on Data Science, Technology and Applications*, 2022, pp. 312–319. doi: 10.5220/0011265900003269.

[23] P. D. Moral, S. Nowaczyk, and S. Pashami, "Why Is Multiclass Classification Hard?," *IEEE Access*, vol. 10, pp. 80448–80462, 2022, doi: 10.1109/ACCESS.2022.3192514.

[24] A. Rahali and M. A. Akhloufi, "End-to-End Transformer-Based Models in Textual-Based NLP," *AI*, vol. 4, no. 1, pp. 54–110, Jan. 2023, doi: 10.3390/ai4010004.

[25] 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: doi.org/10.52436/1.jutif.2025.6.3.4585.

[26] D. Z. Abidin, A. Siswanto, C. Saputra, B. Betantiyo, and A. Nehemia Toscany, "Enhancing Fake News Detection on Imbalanced Data Using Resampling Techniques and Classical Machine Learning Models," *J. Tek. Inform. Jutif*, vol. 6, no. 5, pp. 3769–3786, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5177.

[27] V. K. Agbesi *et al.*, "Pre-Trained Transformer-Based Models for Text Classification Using Low-Resourced Ewe Language," *Systems*, vol. 12, no. 1, p. 1, Dec. 2023, doi: 10.3390/systems12010001.

[28] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, p. 4157, May 2022, doi: 10.3390/s22114157.

[29] M. K. Shaik Vadla, M. A. Suresh, and V. K. Viswanathan, "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT," *Algorithms*, vol. 17, no. 2, p. 59, Jan. 2024, doi: 10.3390/a17020059.

[30] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023, doi: 10.1038/s42256-023-00626-4.

[31] M. M. Krell, M. Kosec, S. P. Perez, and A. Fitzgibbon, "Efficient Sequence Packing without Cross-contamination: Accelerating Large Language Models without Impacting Performance," Oct. 05, 2022, *arXiv*: arXiv:2107.02027. doi: 10.48550/arXiv.2107.02027.

[32] S. Ramakrishnan and L. D. Dhinesh Babu, "Improving Multi-Label Emotion Classification on Imbalanced Social Media Data With BERT and Clipped Asymmetric Loss," *IEEE Access*, vol. 13, pp. 60589–60601, 2025, doi: 10.1109/ACCESS.2025.3557091.

[33] M. Rehan, M. S. I. Malik, and M. M. Jamjoom, "Fine-Tuning Transformer Models Using Transfer Learning for Multilingual Threatening Text Identification," *IEEE Access*, vol. 11, pp. 106503–106515, 2023, doi: 10.1109/ACCESS.2023.3320062.

[34] R. Pan, J. A. García-Díaz, F. Garcia-Sanchez, and R. Valencia-García, "Evaluation of transformer models for financial targeted sentiment analysis in Spanish," *PeerJ Comput. Sci.*, vol. 9, p. e1377, May 2023, doi: 10.7717/peerj-cs.1377.

Vol. 7, No. 2, April 2026, Page. 1964-1980
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5350

[35] M. A. Shah, M. J. Iqbal, N. Noreen, and I. Ahmed, "An Automated Text Document Classification Framework using BERT," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140332.

[36] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 99, Dec. 2022, doi: 10.1007/s13278-022-00934-4.

[37] Y. Cao, Z. Sun, L. Li, and W. Mo, "A Study of Sentiment Analysis Algorithms for Agricultural Product Reviews Based on Improved BERT Model," *Symmetry*, vol. 14, no. 8, p. 1604, Aug. 2022, doi: 10.3390/sym14081604.