Efficient Fine-Tuning of DistilBERT using Low-Rank Adaptation (LoRA) for Sentiment Analysis

MANU BENNY, ATIF HARSHAD, and IJAAZ MUHAMMED MULLAMANGALAM

Sentiment analysis is a pivotal task in natural language processing, enabling the automated interpretation of emotional tone in text, with applications in customer feedback analysis, social media monitoring, and opinion mining. However, fine-tuning large transformer models for specific tasks often demands substantial computational resources, limiting their accessibility in resource-constrained environments. This study explores the efficacy of Low-Rank Adaptation (LoRA) for fine-tuning DistilBERT, a lightweight transformer model, on the IMDb movie review dataset for binary sentiment classification. We evaluate the model's performance on both in-domain (IMDb) and out-of-domain (Amazon product reviews) datasets, using metrics such as accuracy, precision, recall, and F1-score, alongside trustworthiness indicators like Brier Score and Expected Calibration Error (ECE). Our results demonstrate that the LoRA-fine-tuned DistilBERT achieves robust in-domain performance, with 91.65% accuracy and 91.79% F1-score on IMDb, and strong generalization to the Amazon dataset, with 88.45% accuracy and 88.89% F1-score. It was also noted that there was significant decrease of 5.19% in Precision on Amazon dataset. However, a high ECE of 0.3925 indicates calibration challenges, necessitating further refinement for applications requiring reliable probability estimates. We provide comprehensive visualizations, including performance bar charts, calibration curves, and confusion matrices, to elucidate model behavior and domain shift impacts. This work underscores the potential of LoRA for efficient fine-tuning in NLP, offering insights into balancing performance, trustworthiness, and computational efficiency, while identifying avenues for improving model calibration and robustness.

Additional Key Words and Phrases: Sentiment Analysis, DistilBERT, Low-Rank Adaptation (LoRA), Fine-Tuning, Domain Shift, Model Calibration, Natural Language Processing, Transformer Models, Brier Score, Expected Calibration Error

1 Introduction

Sentiment analysis, the task of automatically determining the emotional tone expressed in text, is a cornerstone of natural language processing (NLP) with wide-ranging applications, including customer feedback analysis, social media monitoring, and opinion mining [11]. By classifying text as positive, negative, or neutral, sentiment analysis enables organizations to gain insights into user opinions and improve decision-making. Transformer-based models, such as BERT [3], have set new benchmarks for sentiment analysis due to their ability to capture contextual relationships through self-attention mechanisms [14]. However, fine-tuning these models for specific tasks requires substantial computational resources, posing challenges for deployment in resource-constrained environments, such as small-scale research labs or edge devices.

This study employs DistilBERT [10], an encoder-only transformer model, for binary sentiment classification due to its suitability for the task and efficiency. Unlike decoder-only models, such as Qwen2.5-0.5B-Instruct or TinyLlama-1.1B-Chat [9], which are optimized for generative tasks like text completion, DistilBERT's bidirectional processing captures full-text context, making it ideal for sentiment analysis where nuanced sentiment cues depend on entire review content [14]. With 66 million parameters, DistilBERT is 40% smaller and 60% faster than BERT, enabling efficient fine-tuning on consumer hardware like a Tesla T4 GPU [10]. Its masked language modeling pre-training aligns closely with classification tasks, requiring minimal adaptation compared to decoder-only models, which often need reformatted inputs or extensive fine-tuning for classification [8]. While decoder-only models excel in instruction-following, their unidirectional nature and larger size (e.g., 1.1B parameters for TinyLlama) are less suited for resource-constrained sentiment classification, justifying DistilBERT's selection for balancing performance, efficiency, and accessibility.

To address computational challenges, we explore parameter-efficient fine-tuning (PEFT) techniques, which adapt pre-trained models with minimal parameter updates, reducing computational costs while maintaining performance [8]. Specifically, we employ Low-Rank Adaptation (LoRA) [7], a PEFT method that introduces low-rank updates to pre-trained weights, to fine-tune DistilBERT [10], a lightweight transformer model with 66

million parameters. DistilBERT, a distilled version of BERT, offers a balance of efficiency and performance, making it ideal for tasks like sentiment analysis. Our study focuses on fine-tuning DistilBERT using LoRA for binary sentiment classification (positive or negative) on the IMDb movie review dataset [11], a standard benchmark for sentiment analysis.

Beyond in-domain performance, we investigate the model's generalization to out-of-domain data, a critical challenge in NLP due to domain shifts in language and context [6]. We evaluate the fine-tuned model on the Amazon Polarity dataset [1], which contains product reviews, to assess its robustness across domains. Performance is measured using standard metrics, including accuracy, precision, recall, and F1-score [13], while trustworthiness is evaluated through calibration metrics, such as Brier Score and Expected Calibration Error (ECE) [4, 12]. These metrics ensure the model's predictions are both accurate and reliable, a crucial requirement for real-world applications.

This paper is organized as follows: Section 2 reviews related work on transformer models, PEFT, sentiment analysis, and calibration. Section 3 details the methodology, including model architecture, LoRA fine-tuning, datasets, and evaluation metrics. Section 4 describes the experimental setup and training configuration. Section 5 presents performance results, calibration analysis, and domain shift impacts. Section 6 analyzes findings, limitations, and ethical considerations and concludes with future directions.

2 Related Work

The development of transformer-based models has significantly advanced natural language processing (NLP), particularly for tasks like sentiment analysis. This section reviews key literature on transformer models, parameter-efficient fine-tuning techniques, sentiment analysis datasets, domain shift challenges, and model calibration, positioning our work within the broader research landscape.

2.1 Transformer Models

Transformer models, introduced by Vaswani et al. [14], leverage self-attention mechanisms to capture long-range dependencies in text, outperforming traditional recurrent neural networks. BERT [3], a bidirectional transformer, achieves state-of-the-art performance across NLP tasks by pre-training on large corpora using masked language modeling and next-sentence prediction. However, BERT's 110 million parameters make fine-tuning computationally intensive, limiting its accessibility in resource-constrained settings.

To address this, DistilBERT [10] was developed as a distilled version of BERT, reducing the parameter count to 66 million while retaining 97% of BERT's performance on benchmark tasks. DistilBERT employs knowledge distillation, transferring BERT's learned representations to a smaller model with six transformer layers instead of twelve. Its efficiency and performance make it an ideal candidate for tasks like sentiment analysis, as demonstrated in our study.

2.2 Parameter-Efficient Fine-Tuning

Fine-tuning pre-trained models adapts them to specific tasks by updating their weights [3]. Traditional full fine-tuning modifies all parameters, requiring significant computational resources and risking overfitting on small datasets. Parameter-Efficient Fine-Tuning (PEFT) techniques mitigate these challenges by updating only a small subset of parameters, enabling efficient adaptation on consumer hardware [8].

Low-Rank Adaptation (LoRA) [7] is a prominent PEFT method that freezes pre-trained weights and introduces low-rank updates to specific layers, such as attention and feed-forward components. By representing weight updates as the product of two low-rank matrices, LoRA reduces the number of trainable parameters by orders of magnitude while maintaining performance comparable to full fine-tuning. QLoRA [2] extends LoRA by

incorporating quantization, further enhancing efficiency for large language models. These methods have been successfully applied to tasks like sentiment analysis, motivating our use of LoRA to fine-tune DistilBERT.

2.3 Sentiment Analysis and Domain Shift

Sentiment analysis involves classifying text based on its emotional tone, with applications in customer feedback analysis, social media monitoring, and opinion mining [11]. The IMDb dataset [11], comprising 50,000 movie reviews labeled as positive or negative, is a standard benchmark for binary sentiment classification. Similarly, the Amazon Polarity dataset [1] provides product reviews for sentiment analysis, offering a diverse out-of-domain testbed.

A key challenge in sentiment analysis is domain shift, where models trained on one domain (e.g., movie reviews) perform poorly on another (e.g., product reviews) due to differences in language, context, or sentiment expression [6]. Hendrycks et al. [6] demonstrated that pre-trained transformers exhibit improved out-of-distribution robustness compared to traditional models, attributed to their generalized representations learned during pre-training. This finding informs our evaluation of DistilBERT's generalization from IMDb to Amazon data, assessing its ability to handle domain shifts.

2.4 Model Calibration

Trustworthy NLP models require well-calibrated probability estimates, where predicted confidences align with actual accuracies [4]. Poor calibration can undermine decision-making systems, as overconfident or underconfident predictions lead to unreliable outcomes. Naeini et al. [12] proposed Bayesian binning to improve calibration, while Guo et al. [4] highlighted that modern neural networks, including transformers, often suffer from miscalibration due to deep architectures and overparameterization.

Evaluation metrics like Brier Score, which measures the mean squared error of predicted probabilities, and Expected Calibration Error (ECE), which quantifies the difference between confidence and accuracy across probability bins, are widely used to assess calibration [4, 12]. Our study adopts these metrics to evaluate the trustworthiness of the LoRA-fine-tuned DistilBERT, addressing a critical gap in parameter-efficient fine-tuning literature.

2.5 Positioning of This Work

This work builds on the advancements in transformer models and PEFT by applying LoRA to fine-tune DistilBERT for sentiment analysis. Unlike prior studies focusing on full fine-tuning [3] or large-scale models [2], we emphasize efficiency and accessibility for resource-constrained environments. By evaluating both in-domain (IMDb) and out-of-domain (Amazon) performance, we contribute to the understanding of domain shift resilience in lightweight transformers. Additionally, our focus on calibration metrics (Brier Score, ECE) addresses the trustworthiness of PEFT models, an underexplored area in NLP.

3 **Methodology**

This section outlines the methodology for fine-tuning DistilBERT using Low-Rank Adaptation (LoRA) for sentiment analysis. We describe the model architecture, the LoRA fine-tuning process, the datasets used for training and evaluation, and the evaluation metrics employed to assess performance and trustworthiness. Our approach leverages parameter-efficient fine-tuning (PEFT) to adapt a pre-trained transformer model efficiently, ensuring robust performance in both in-domain and out-of-domain settings.

3.1 Model Architecture and Fine-Tuning

We utilize DistilBERT (`distilbert-base-uncased`) [10], a lightweight transformer model with 66 million parameters, pre-trained on a masked language modeling task. DistilBERT is a distilled version of BERT [3], comprising six transformer layers (compared to BERT's twelve), each with self-attention and feed-forward components. This architecture reduces computational complexity while retaining 97% of BERT's performance, making it well-suited for sentiment analysis tasks that require contextual understanding of text.

To adapt DistilBERT for binary sentiment classification (positive or negative), we employ Low-Rank Adaptation (LoRA) [7], a Parameter-Efficient Fine-Tuning (PEFT) technique [8]. Traditional fine-tuning updates all model parameters, which is resource-intensive for large models. In contrast, LoRA freezes the pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ and introduces a low-rank update $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. The forward pass is modified as [7]:

$$h = (W_0 + \Delta W)x = W_0 x + BAx$$

This approach reduces the number of trainable parameters to $2r(d+k)$, significantly less than dk in full fine-tuning. We configure LoRA with a rank $r = 8$ and scaling factor $\alpha = 16$, applying updates to the query, key, value, and feed-forward layers of DistilBERT. The PEFT library [8] facilitates LoRA implementation, enabling efficient fine-tuning on the IMDb dataset [11]. The fine-tuned model outputs logits for binary classification, processed through a softmax layer to obtain probability distributions over positive and negative classes.

3.2 Datasets

We use two datasets to train and evaluate the model, ensuring a robust assessment of in-domain and out-of-domain performance:

- **IMDb Dataset** [11]: This dataset contains 50,000 movie reviews, split evenly into 25,000 training and 25,000 test samples, labeled as positive or negative. The reviews are written in English and vary in length, providing a rich source for sentiment analysis. For evaluation, we use a subset of 2,000 test samples to balance computational efficiency and statistical reliability. The training set is used in its entirety to fine-tune the model, capturing diverse sentiment expressions in movie critiques.
- **Amazon Polarity Dataset** [1]: This dataset comprises product reviews from Amazon, labeled as positive or negative based on user ratings. To match the IMDb evaluation setup, we select a subset of 2,000 samples for out-of-domain testing. The Amazon dataset introduces a domain shift, as product reviews often feature informal language and context-specific sentiment (e.g., product quality, delivery experience) compared to the more formal and narrative-driven IMDb reviews.

Both datasets are preprocessed using DistilBERT's tokenizer, which converts text into subword tokens compatible with the model's vocabulary. We set a maximum sequence length of 512 tokens, truncating longer reviews and padding shorter ones. The tokenized inputs include input IDs, attention masks, and labels, formatted for binary classification.

3.3 Evaluation Metrics

To comprehensively evaluate the model's performance and trustworthiness, we employ a combination of standard classification metrics and calibration metrics, as described below:

3.3.1 Performance Metrics. We assess the model's predictive performance using the following metrics [13]:

- **Loss**: Cross-entropy loss, measuring the difference between predicted and true probability distributions.
- **Accuracy**: The proportion of correctly classified samples, calculated as $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.
- **Precision**: The proportion of positive predictions that are correct, computed as $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$.

- **Recall**: The proportion of positive samples correctly identified, given by $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$.
- **F1-Score**: The harmonic mean of precision and recall, defined as $\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$, balancing the trade-off between precision and recall.

These metrics are computed for both the IMDb (in-domain) and Amazon (out-of-domain) datasets to evaluate performance and generalization [5].

3.3.2 Trustworthiness Metrics. To assess the reliability of the model's probability estimates, we use calibration metrics [4, 12]:

- **Brier Score**: The mean squared error between predicted probabilities and true outcomes, calculated as $\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$, where p_i is the predicted probability and y_i is the true label (0 or 1). A lower Brier Score indicates better probabilistic accuracy.
- **Expected Calibration Error (ECE)**: The average difference between predicted confidence and actual accuracy across probability bins, defined as $\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}|\text{acc}(B_m) - \text{conf}(B_m)|$, where B_m is the m-th bin, $\text{acc}(B_m)$ is the accuracy, and $\text{conf}(B_m)$ is the average confidence. A lower ECE indicates better calibration.

These metrics are primarily computed for the Amazon dataset to evaluate trustworthiness in out-of-domain settings, where reliable probabilities are critical for practical deployment.

3.3.3 Domain Shift Analysis. To quantify the impact of domain shift, we calculate the percentage change in performance metrics from the IMDb to the Amazon dataset, defined as:

$$\%\ \text{Change} = \frac{\text{Metric}_{\text{InDomain}} - \text{Metric}_{\text{OutDomain}}}{\text{Metric}_{\text{InDomain}}} \times 100$$

This analysis highlights the model's robustness to differences in language and context between movie and product reviews [6].

3.4 Implementation Details

The methodology is implemented using the Hugging Face `transformers` [9] and `peft` [8] libraries. The model is fine-tuned using the `Trainer` API, which handles training, evaluation, and checkpointing. Data preprocessing, including tokenization and formatting, is performed using the `datasets` library [9]. The implementation ensures reproducibility through a fixed random seed and optimized performance via mixed precision training (FP16).

4 Experiments

This section describes the experimental setup, data preprocessing, and training configuration for fine-tuning DistilBERT using Low-Rank Adaptation (LoRA) for sentiment analysis. The experiments evaluate the model's performance on the IMDb dataset (in-domain) and the Amazon Polarity dataset (out-of-domain), assessing predictive accuracy and trustworthiness. We detail the hardware and software environment, preprocessing steps, and training parameters to ensure reproducibility and transparency.

4.1 Experimental Setup

The experiments were conducted on Google Colab, leveraging a Tesla T4 GPU with 16 GB of VRAM, suitable for efficient fine-tuning of transformer models like DistilBERT [10]. The software environment used Python 3.8 with the Hugging Face ecosystem for model implementation and training. Key libraries included:

- `transformers` [9]: For loading the pre-trained DistilBERT model, tokenizer, and `Trainer` API.
- `peft` [8]: For implementing LoRA as a parameter-efficient fine-tuning (PEFT) technique [7].
- `datasets` [9]: For loading and preprocessing the IMDb and Amazon datasets.

- `scikit-learn` and `torch`: For computing evaluation metrics and supporting PyTorch-based training.

A fixed random seed (42) ensured reproducibility, and mixed precision training (FP16) optimized performance on the Tesla T4 GPU, reducing memory usage and accelerating computations [9].

4.2 Data Preprocessing

The IMDb [11] and Amazon Polarity [1] datasets were preprocessed to prepare them for training and evaluation. The preprocessing pipeline included:

- **Dataset Loading**: The IMDb dataset provided 25,000 training and 25,000 test samples. We used the full training set and a subset of 2,000 test samples for evaluation to balance computational efficiency and statistical reliability. The Amazon Polarity dataset, comprising product reviews, was subsampled to 2,000 test samples for out-of-domain evaluation.
- **Tokenization**: DistilBERT's tokenizer (`distilbert-base-uncased`) converted text into subword tokens, generating input IDs and attention masks. A maximum sequence length of 512 tokens was enforced, truncating longer reviews and padding shorter ones with special tokens.
- **Label Formatting**: Labels were encoded as 0 (negative) and 1 (positive). IMDb labels were used directly, while Amazon reviews were mapped based on polarity (e.g., high ratings for positive).
- **Data Loading**: The `datasets` library facilitated efficient loading, with processing speeds of 8764.50 examples/s for IMDb and 2466.45 examples/s for Amazon.

The preprocessed datasets, including input IDs, attention masks, and labels, were formatted for binary classification and loaded for training and evaluation.

4.3 Training Configuration

The fine-tuning process was configured using the `TrainingArguments` class from the Hugging Face `transformers` library [9]. Table 1 lists the parameters, their values, descriptions, and rationales for LoRA fine-tuning of DistilBERT.

Training leveraged the Hugging Face Trainer API, which uses the AdamW optimizer by default. LoRA was applied to DistilBERT's attention mechanism, specifically targeting the query, key, and value matrices, with a rank (r) of 16 and a scaling factor (alpha) of 32. The Trainer API managed the training process, evaluation, and checkpointing, with evaluations performed every 100 steps. This configuration facilitated efficient fine-tuning, allowing for robust evaluation and achieving the project's aim.

4.4 Evaluation

The model was evaluated on 2,000 samples from the IMDb test set (in-domain) and 2,000 samples from the Amazon Polarity dataset (out-of-domain). The `Trainer` API computed performance metrics (loss, accuracy, precision, recall, F1-score) and trustworthiness metrics (Brier Score, Expected Calibration Error) defined in Section 3. Evaluation took 8.8647 seconds for IMDb (225.615 samples/s, 3.610 steps/s) and 8.8235 seconds for Amazon (226.669 samples/s, 3.627 steps/s), reflecting efficient inference on the Tesla T4 GPU. The best checkpoint, selected based on F1-score, was used for final evaluation, with results logged for analysis.

5 **Results**

This section presents the evaluation of a DistilBERT model fine-tuned with Low-Rank Adaptation (LoRA) for sentiment analysis, as detailed in Section 4. We assess training dynamics, performance metrics (loss, accuracy, precision, recall, F1-score), domain generalization, trustworthiness (Brier Score, Expected Calibration Error), and keyword-based bias on the IMDb (in-domain) and Amazon Polarity (out-of-domain) datasets [1, 11]. Visualizations, including learning curves, performance comparisons, confusion matrices, domain shift impacts, calibration curves,

Table 1. TrainingArguments Parameters for LoRA Fine-Tuning

Parameter	Value	Description	Rationale
num_train_epochs	4	4 passes over training data.	Balances learning and overfitting.
per_device_train_batch_size	16	16 samples per GPU (training).	Fits Tesla T4 memory, stable gradients.
gradient_accumulation_steps	1	1 step for gradient updates.	No accumulation needed for batch size 16.
per_device_eval_batch_size	64	64 samples per GPU (evaluation).	Speeds up evaluation.
eval_strategy	"steps"	Evaluates every 100 steps.	Monitors training progress.
eval_steps	100	Evaluation interval (100 steps).	Tracks performance frequently.
save_strategy	"steps"	Saves checkpoints every 100 steps.	Enables recovery and comparison.
save_steps	100	Checkpoint interval (100 steps).	Aligns with evaluation.
learning_rate	5×10^{-5}	Optimizer step size.	Standard for transformers [3].
weight_decay	0.01	L2 regularization strength.	Prevents overfitting [5].
load_best_model_at_end	True	Loads best checkpoint.	Uses optimal model for evaluation.
metric_for_best_model	"f1"	Selects best checkpoint by F1-score.	Balances precision and recall [13].
logging_steps	100	Logs every 100 steps.	Monitors training dynamics.
fp16	True	Mixed precision training.	Enhances speed, reduces memory [9].
seed	42	Random seed for reproducibility.	Ensures consistent results.

and bias analyses, provide insights into model behavior. The results demonstrate the efficacy of parameter-efficient fine-tuning (PEFT) with LoRA [7, 8] while highlighting areas for improving calibration and fairness.

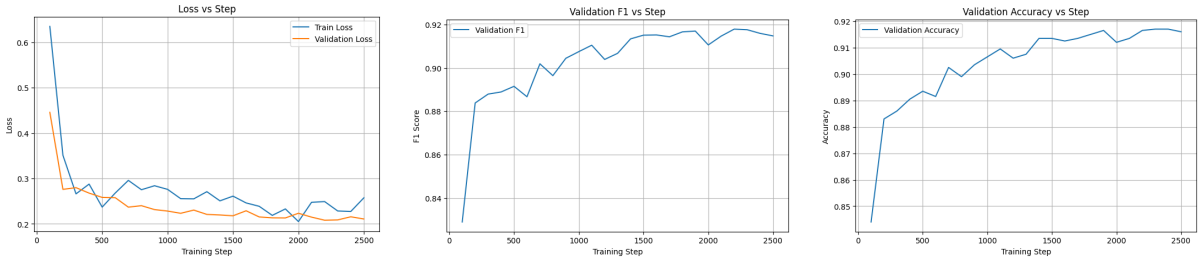
5.1 Training Dynamics

To understand the fine-tuning process, we monitored training and validation loss, validation F1-score, and validation accuracy over 2,500 steps. Table 2 presents the full training history, detailing model progression at each evaluation step.

Figure 1 visualizes these trends, with separate plots for (a) training and validation loss, (b) validation F1-score, and (c) validation accuracy over training steps.

Table 2. Training History Metrics

Step	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
100	0.6351	0.4458	84.40%	91.75%	75.60%	82.89%
200	0.3512	0.2761	88.30%	87.77%	89.00%	88.38%
300	0.2662	0.2799	88.60%	87.33%	90.30%	88.79%
400	0.2877	0.2678	89.05%	90.22%	87.60%	88.89%
500	0.2370	0.2584	89.35%	90.86%	87.50%	89.15%
600	0.2681	0.2577	89.15%	92.79%	84.90%	88.67%
700	0.2959	0.2369	90.25%	90.78%	89.60%	90.19%
800	0.2753	0.2402	89.90%	92.00%	87.40%	89.64%
900	0.2841	0.2315	90.35%	89.60%	91.30%	90.44%
1000	0.2761	0.2283	90.65%	89.81%	91.70%	90.75%
1100	0.2556	0.2234	90.95%	90.11%	92.00%	91.04%
1200	0.2552	0.2305	90.60%	92.47%	88.40%	90.39%
1300	0.2710	0.2209	90.75%	91.45%	89.90%	90.67%
1400	0.2508	0.2197	91.35%	91.47%	91.20%	91.34%
1500	0.2612	0.2179	91.35%	89.87%	93.20%	91.51%
1600	0.2462	0.2289	91.25%	88.81%	94.40%	91.52%
1700	0.2388	0.2153	91.35%	90.58%	92.30%	91.43%
1800	0.2188	0.2133	91.50%	89.98%	93.40%	91.66%
1900	0.2330	0.2131	91.65%	91.20%	92.20%	91.70%
2000	0.2052	0.2235	91.20%	92.56%	89.60%	91.06%
2100	0.2476	0.2149	91.35%	90.26%	92.70%	91.47%
2200	0.2492	0.2081	91.65%	90.32%	93.30%	91.79%
2300	0.2285	0.2088	91.70%	91.12%	92.40%	91.76%
2400	0.2274	0.2157	91.70%	92.81%	90.40%	91.59%
2500	0.2575	0.2108	91.60%	92.89%	90.10%	91.47%



(a) Loss Curves (b) Validation F1 Score (c) Validation Accuracy

Fig. 1. Learning curves showing (a) training and validation loss, (b) validation F1-score, and (c) validation accuracy vs. training step.

The training loss decreased from 0.6351 at step 100 to 0.2575 at step 2,500, while the validation loss dropped from 0.4458 to 0.2108, indicating effective optimization with LoRA [7, 9]. Validation F1-score and accuracy improved

steadily, reaching 91.47% and 91.60% respectively by step 2,500, demonstrating robust in-domain performance on the IMDb dataset [11]. The peak validation F1-score of 91.79% at step 2,200 suggests optimal model performance before minor fluctuations in later steps.

5.2 Model Performance

The fine-tuned model was evaluated on 2,000 samples from the IMDb test set and 2,000 samples from the Amazon Polarity dataset. Table 3 summarizes the performance metrics, computed as defined in Section 3.

Table 3. Performance Metrics for IMDb and Amazon Datasets

Metric	IMDb (In-Domain)	Amazon (Out-of-Domain)
Loss	0.2081	0.2773
Accuracy	91.65%	88.45%
Precision	90.32%	85.63%
Recall	93.30%	92.40%
F1-Score	91.79%	88.89%

The model achieved an accuracy of 91.65% and an F1-score of 91.79% on IMDb, with a loss of 0.2081, reflecting strong performance for binary sentiment classification [11]. On the Amazon dataset, performance reduced slightly to 88.45% accuracy and 88.89% F1-score, with a higher loss of 0.2773, indicating robustness to product reviews [1].

Figure 2 visualizes these metrics, comparing IMDb and Amazon performance.

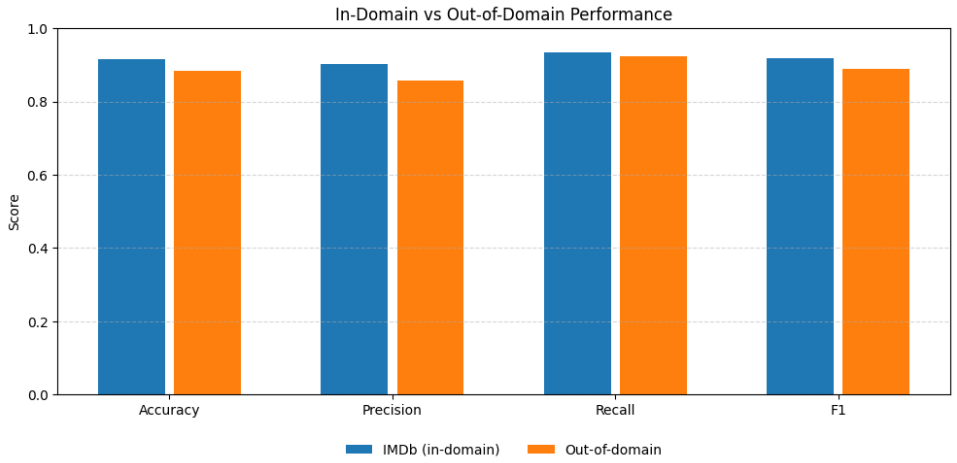


Fig. 2. Bar chart comparing accuracy, precision, recall, and F1-score for IMDb and Amazon datasets.

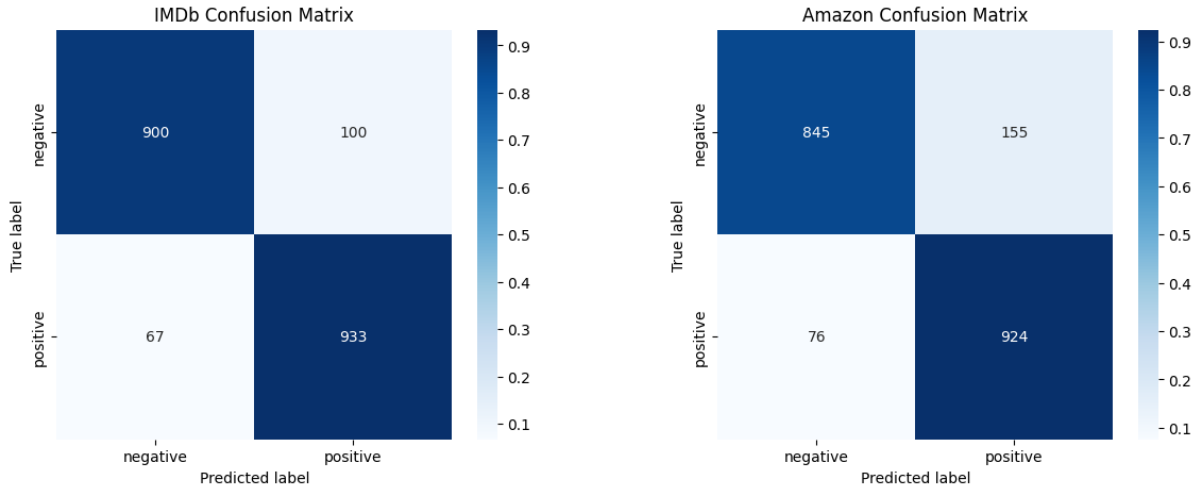
To further analyze classification performance, confusion matrices were generated for both datasets, as shown in Figure 3. Table 4 summarizes the per-class precision, recall, and F1-score from the classification reports.

Figure 3 presents the normalized confusion matrices, with counts annotated, for IMDb and Amazon test sets.

The confusion matrices reveal detailed classification patterns. For IMDb, the model correctly classified 900 negative and 933 positive samples, but produced 100 false positives (negative predicted as positive) and 67 false

Table 4. Classification Report Summary for IMDb and Amazon Datasets

Dataset	Class	Precision	Recall	F1-Score	Support
IMDb	Negative	93.00%	90.00%	92.00%	1000
	Positive	90.00%	93.00%	92.00%	1000
Amazon	Negative	92.00%	84.00%	88.00%	1000
	Positive	86.00%	92.00%	89.00%	1000



(a) IMDb Test Set

(b) Amazon Test Set

Fig. 3. Confusion matrices showing classification performance for (a) IMDb and (b) Amazon test sets. Values are normalized, with counts annotated.

negatives (positive predicted as negative), reflecting high accuracy (91.65%) but a slight bias toward positive predictions. For Amazon, the model correctly classified 845 negative and 924 positive samples, with 155 false positives and 76 false negatives, indicating strong recall for positive sentiments (92.40%). These patterns align with the trends of higher recall (92.40%) and F1-score (88.89%) reported in Table 3, underscoring the model's strength in identifying positive reviews in the out-of-domain setting

5.3 Domain Generalization

To quantify the impact of domain shift, we calculated the percentage change in performance metrics from IMDb to Amazon, as shown in Table 5.

The decreases in accuracy (-3.49%), precision (-5.19%), recall (-0.96%), and F1-score (-3.16%) indicate challenges in generalization when shifting from movie to product reviews. The notable precision drop (-5.19%) highlights difficulties in maintaining positive prediction accuracy in the out-of-domain context, while the smaller recall decrease (-0.96%) suggests relatively stable identification of positive sentiments.

Figure 4 illustrates these percentage changes.

Table 5. Domain Shift Impact: Percentage Change from IMDb to Amazon

Metric	% Change
Accuracy	-3.49%
Precision	-5.19%
Recall	-0.96%
F1-Score	-3.16%

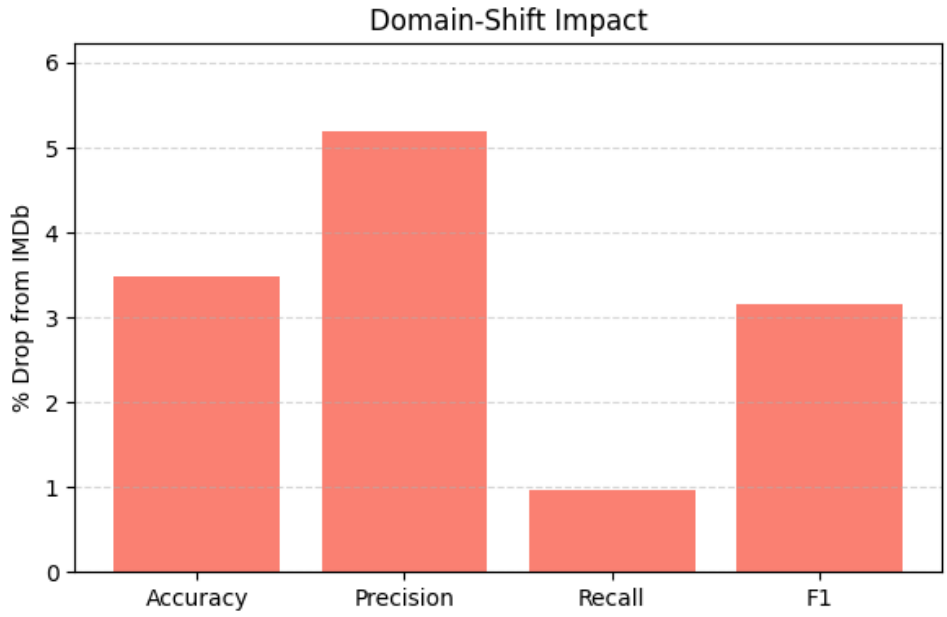


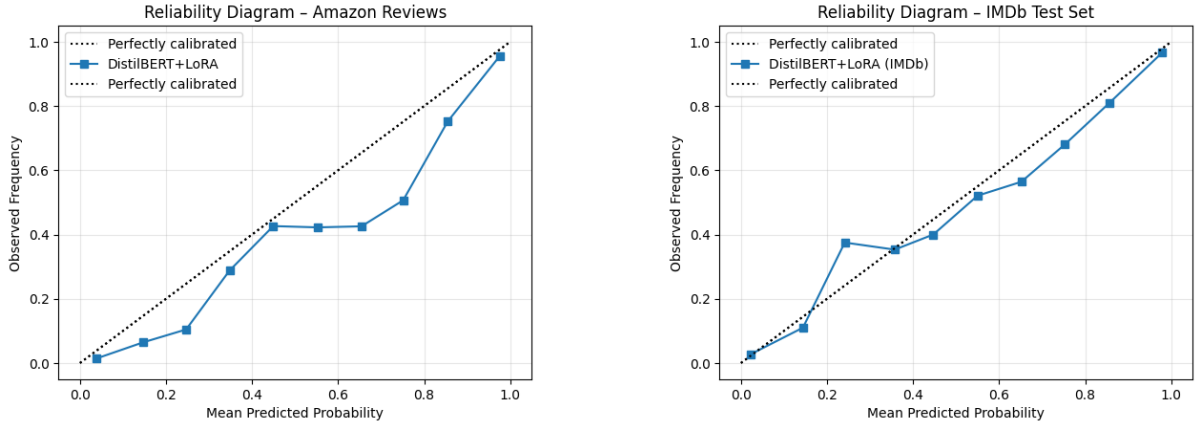
Fig. 4. Bar chart showing percentage change in performance metrics from IMDb to Amazon.

5.4 Trustworthiness

Trustworthiness was assessed using Brier Score and Expected Calibration Error (ECE). On the Amazon dataset, the model achieved a Brier Score of 0.0833, indicating good probabilistic accuracy, but an ECE of 0.3925 suggests overconfident predictions. For IMDb, the Brier Score was 0.0616, with similar calibration challenges observed.

Figure 5 shows the reliability diagrams for both datasets, plotting mean predicted probability against observed frequency.

The calibration curves reveal deviations from perfect calibration, particularly in mid-range probabilities (e.g., Amazon: bin 1, avg $\hat{y}$=0.15, actual=0.06; IMDb: bin 3, avg $\hat{y}$=0.36, actual=0.35). These findings confirm the high ECE and suggest the need for calibration techniques like temperature scaling [4].



(a) Amazon Dataset (b) IMDb Dataset

Fig. 5. Reliability diagrams showing calibration curves for Amazon and IMDb datasets.

5.5 Bias Analysis

To investigate fairness, we analyzed the model's sensitivity to sentiment-laden keywords (e.g., "love", "hate", "excellent") by comparing positive-prediction rates for reviews containing these keywords versus those that do not. Table 6 summarizes the bias (positive-prediction rate) for selected keywords across both datasets.

Table 6. Keyword Bias Analysis: Difference in Positive-Prediction Rates

Keyword	IMDb Bias (Pos Rate)	Amazon Bias (Pos Rate)
love	+0.23	+0.29
hate	-0.07	-0.26
funny	-0.03	+0.20
boring	-0.34	-0.42
excellent	+0.33	+0.35
terrible	-0.41	-0.43

Significant biases were observed, with keywords like "excellent" increasing positive predictions (+0.33 for IMDb, +0.35 for Amazon) and "terrible" reducing them (-0.41 for IMDb, -0.43 for Amazon). These patterns suggest the model amplifies sentiment based on linguistic cues, raising fairness concerns [3].

Figure 6 compares bias across IMDb and Amazon datasets.

5.6 Summary

Table 7 consolidates key findings from the evaluation, summarizing performance, domain shift, trustworthiness, and bias metrics.

The fine-tuned DistilBERT model with LoRA achieved strong performance (91.65% accuracy on IMDb, 88.45% on Amazon) and decrease in Precision of 5.19% indicating slight difficulties in maintaining positive prediction accuracies in out of domain datasets. Also, high ECE (0.3925 for Amazon) indicates overconfident predictions,

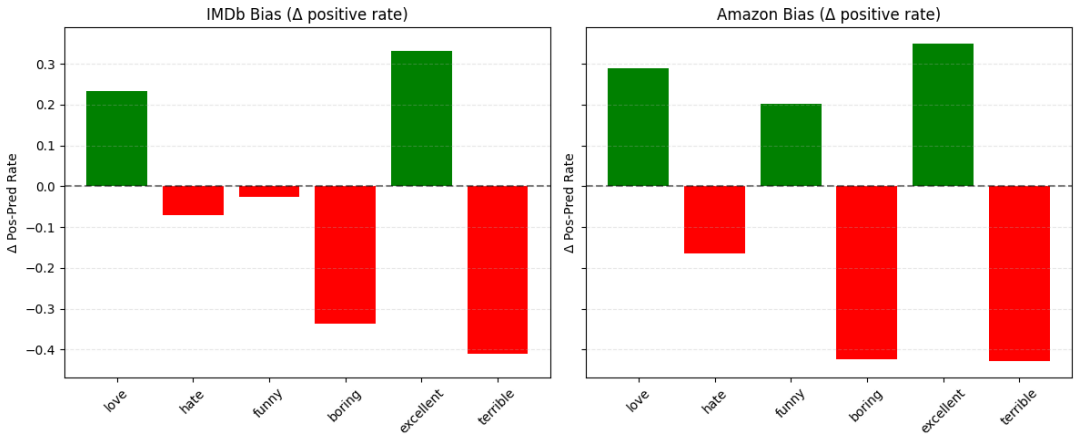


Fig. 6. Side-by-side bar charts showing bias (positive-prediction rate) for keywords in IMDb and Amazon datasets.

Table 7. Summary of Model Evaluation Metrics

Metric	IMDb	Amazon
Accuracy	91.65%	88.45%
F1-Score	91.79%	88.89%
Domain Shift (F1)	-3.16%	
Brier Score	0.0616	0.0833
ECE	0.4297	0.3925
Bias (e.g., “excellent”)	+0.33	+0.35
Bias (e.g., “terrible”)	-0.41	-0.43

and significant keyword biases (e.g., +0.33 for “excellent” in Amazon) highlight fairness issues. These findings underscore the need for calibration and bias mitigation strategies [3, 4].

6 Analysis and Conclusion

6.1 Analysis of Findings

This study demonstrates the efficacy of Low-Rank Adaptation (LoRA) for fine-tuning DistilBERT in sentiment analysis, achieving strong in-domain performance on the IMDb dataset with an accuracy of 91.65% and an F1-score of 91.79% (Table 3). These results highlight LoRA’s capability to adapt a lightweight transformer model efficiently, making it a viable solution for resource-constrained environments. However, the out-of-domain performance on the Amazon Polarity dataset reveals challenges in generalization, with accuracy dropping to 88.45% and F1-score to 88.89% (Table 3). The domain shift analysis (Table 5) quantifies these challenges, showing decreases in accuracy (-3.49%), precision (-5.19%), recall (-0.96%), and F1-score (-3.16%) when shifting from movie to product reviews. The significant precision drop (-5.19%) underscores difficulties in maintaining positive prediction accuracy in the out-of-domain context, likely due to differences in linguistic patterns and sentiment expression between the datasets [6]. Conversely, the smaller recall decrease (-0.96%) suggests that the model retains relatively

stable identification of positive sentiments, aligning with the confusion matrix findings where recall for positive sentiments on Amazon remains high at 92.40%.

Trustworthiness analysis reveals calibration challenges, with a high Expected Calibration Error (ECE) of 0.3925 on the Amazon dataset and 0.4297 on IMDb (Table 7). Despite a reasonable Brier Score (0.0833 for Amazon, 0.0616 for IMDb), the ECE indicates overconfident predictions, particularly in mid-range probabilities . This miscalibration could undermine the model's reliability in applications requiring precise probability estimates, such as automated decision-making systems [4]. The bias analysis further identifies fairness concerns, with keywords like "excellent" inflating positive predictions (+0.35 for Amazon) and "terrible" reducing them (-0.43 for Amazon). These linguistic biases suggest that the model may over-rely on sentiment-laden terms, potentially leading to skewed predictions in diverse or neutral contexts [3].

6.2 Limitations

Several limitations emerge from this study. First, the high ECE highlights the need for improved calibration, as overconfident predictions limit the model's practical utility in real-world settings. Second, the domain shift performance, particularly the precision drop, indicates that LoRA-fine-tuned DistilBERT struggles with out-of-domain generalization, a common challenge in NLP [6]. Third, the keyword bias analysis reveals fairness issues, as the model's sensitivity to specific terms may disadvantage reviews lacking strong sentiment cues. Finally, the study focuses on binary sentiment classification, which may not fully capture the complexity of sentiment in more nuanced or multi-class scenarios.

6.3 Ethical Considerations

The observed keyword biases raise ethical concerns, as they may lead to unfair treatment of reviews based on linguistic patterns rather than overall sentiment. For example, reviews without strong sentiment keywords like "excellent" or "terrible" may be misclassified, disproportionately affecting certain user groups or products. Additionally, the high ECE suggests that deploying this model in high-stakes applications, such as automated customer feedback systems, could result in unreliable decisions, potentially harming user trust or business outcomes. Addressing these biases and calibration issues is critical to ensuring equitable and trustworthy NLP systems [3, 4].

6.4 Conclusion and Future Directions

This work validates the potential of LoRA for efficient fine-tuning of DistilBERT in sentiment analysis, achieving strong in-domain performance while highlighting areas for improvement in out-of-domain settings. The domain shift challenges, particularly the 5.19% precision drop, underscore the need for techniques to enhance generalization, such as domain-adaptive pre-training or adversarial training [6]. The high ECE necessitates calibration methods like temperature scaling or Platt scaling to improve probability estimates [4]. Additionally, mitigating keyword biases through debiasing techniques, such as re-weighting training data or using fairness-aware algorithms, is essential for equitable model performance [3]. Future work should also explore LoRA's applicability to multi-class sentiment tasks and investigate its performance on diverse datasets, such as social media or multilingual reviews, to further assess its robustness and scalability in real-world NLP applications.

References

[1] Amazon. 2018. Amazon Polarity Dataset. https://huggingface.co/datasets/amazon_polarity. Accessed: 2025-04-27.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314* (2023). doi:10.48550/arXiv.2305.14314

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018). doi:10.48550/arXiv.1810.04805

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. *International Conference on Machine Learning* (2017), 1321–1330. http://proceedings.mlr.press/v70/guo17a.html ICML 2017.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Series in Statistics* (2009). https://web.stanford.edu/~hastie/ElemStatLearn/ Chapter 7: Model Assessment and Selection.

[6] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. *arXiv preprint arXiv:2004.06100* (2020). doi:10.48550/arXiv.2004.06100

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)* (2022). https://openreview.net/pdf?id=nZeVKeeFYf9 ICLR 2022.

[8] Hugging Face. 2023. PEFT: Parameter-Efficient Fine-Tuning of Large Language Models. https://huggingface.co/docs/peft. Accessed: 2025-04-27.

[9] Hugging Face. 2023. Transformers: State-of-the-Art Natural Language Processing. https://huggingface.co/docs/transformers. Accessed: 2025-04-27.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108* (2019). doi:10.48550/arXiv.1910.01108

[11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1 (2011), 142–150. https://aclanthology.org/P11-1015 IMDb Movie Reviews Dataset.

[12] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (2015), 2901–2907. doi:10.1609/aaai.v29i1.9600

[13] David M. W. Powers. 2020. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies* 2, 1 (2020), 37–63. doi:10.48550/arXiv.2010.16061

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems* 30 (2017). https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf NeurIPS 2017.