Sentiment Analysis of News Headlines Using QLoRA Fine-Tuned LLaMA-2

Sai Vishruth Valandas, Mohammed Sabith Abul Faiz, Tarun Teja Nallan Chakravertula, Abdullah Mohammed
Department of Computer Science Engineering
University of North Texas, Denton, Texas, USA

***Abstract*—The rapid dissemination of news through digital media underscores the critical importance of accurately identifying the sentiment of news headlines. Misinterpretation or incorrect sentiment classification can significantly impact public perception and decision-making. This study develops an effective sentiment analysis model explicitly tailored for brief, contextually rich news headlines. Utilizing the Quantized Low-Rank Adaptation (QLoRA) approach, we fine-tune the pre-trained LLaMA-2-7B large language model on a modest dataset consisting of 1,950 headlines. The proposed methodology significantly reduces computational resource requirements, making it suitable for environments with limited processing power. Our experimental evaluation reveals an accuracy of 82.7, substantially surpassing conventional baseline methods. Additionally, balanced performance metrics across sentiment categories demonstrate the robustness and reliability of our model. This research highlights efficient fine-tuning techniques, contributing to the democratization of powerful language modeling technologies, and making them accessible for broader research and real-world applications.**

***Index Terms*—Sentiment Analysis, LLaMA-2, QLoRA, Low-Rank Adaptation, 4-bit Quantization, Fine-Tuning, Natural Language Processing, Transformer Models, News Headlines, Resource-Efficient NLP, Model Compression, Deep Learning, Gradient Accumulation, AdamW Optimizer.**

I. Introduction

In the contemporary digital landscape, news headlines play a pivotal role in shaping public opinion, influencing societal attitudes, and guiding public discourse on critical issues. Given their significant societal influence, understanding the sentiment behind news headlines accurately is essential. However, news headlines are typically brief and concise, often designed to capture attention swiftly, leaving minimal room for detailed context. This brevity frequently results in ambiguity and challenges traditional natural language processing (NLP) techniques aimed at sentiment analysis.

Sentiment analysis—the computational task of identifying and categorizing opinions expressed in textual data—has become increasingly crucial for analyzing media content, particularly in the fast-paced world of digital news. Accurate sentiment classification assists in various critical tasks such as content moderation, detection and management of misinformation, understanding public opinion trends, and analyzing media biases. However, conventional sentiment analysis methods and machine learning models often fail to capture the nuanced contextual meanings that short texts, such as news headlines, inherently contain.

Recent advancements in NLP have primarily leveraged transformer-based models, such as BERT, RoBERTa, and GPT-3, which have shown promising results in many language understanding tasks. Despite their success, these models typically demand extensive computational resources, particularly when fine-tuning on specific tasks and smaller datasets. This limitation presents a significant barrier for researchers and practitioners operating in resource-constrained environments, thereby restricting widespread access to advanced NLP capabilities.

To overcome these challenges, our study introduces a novel approach by fine-tuning the pre-trained LLaMA-2-7B large language model using Quantized Low-Rank Adaptation (QLoRA). QLoRA is a recent fine-tuning technique designed explicitly for large models, significantly reducing computational demands through weight quantization and low-rank adaptation. By implementing QLoRA, we make sophisticated NLP techniques accessible and practical even on limited hardware.

In this paper, we present a detailed account of our approach, highlighting our contributions to low-resource NLP methodologies and providing extensive empirical evidence through rigorous experimentation. Our proposed model achieves a remarkable accuracy of 82.7, demonstrating significant improvement over baseline methods and underscoring the practicality of efficient fine-tuning strategies. Thus, this study contributes both methodologically and practically, facilitating broader access to state-of-the-art NLP capabilities and paving the way for further research in resource-efficient sentiment analysis.

II. Objective

The primary goal of this research is to develop a highly accurate and efficient sentiment classifier specifically designed for news headlines. The model must accurately classify headlines into positive, negative, or neutral categories while addressing the unique challenges posed by the brevity and context-rich nature of headline texts. Achieving this objective requires the integration of advanced NLP techniques capable of extracting and understanding nuanced contextual information embedded within concise textual formats. Additionally, the classifier should perform reliably with limited computational resources, making it practical for broader adoption across various platforms, including smaller-scale enterprises and academic research environments. This study further aims

to demonstrate the potential of fine-tuning large-scale pre-trained language models through novel methods like Quantized Low-Rank Adaptation (QLoRA), which significantly reduces the computational overhead typically associated with large-scale NLP models. Ultimately, this research endeavors to contribute a practical tool to enhance media monitoring, content moderation, and analytical assessments of public sentiment trends, thus providing a foundation for more robust and accessible sentiment analysis applications.

III. Problem Statement

Sentiment analysis of news headlines presents specific computational challenges due to their inherently brief and often ambiguous nature. Traditional sentiment classification methods frequently depend on substantial textual context to correctly interpret the sentiment; however, headlines typically contain minimal explicit context, leaving significant room for ambiguity and misclassification. As a result, conventional NLP techniques and basic machine learning models often produce suboptimal outcomes, misinterpreting the subtle nuances of headline sentiments. Furthermore, existing sophisticated transformer-based language models, although accurate, require significant computational resources, making their practical application challenging, especially for institutions with limited computing capabilities. The main challenge thus involves creating a model capable of accurately interpreting sentiment from succinct and context-rich news headlines while being resource-efficient enough to be deployed practically across diverse operational settings.

IV. Scope of the Project

This research is carefully scoped around sentiment analysis applied explicitly to a dataset of 1,950 news headlines. The dataset comprises balanced sentiment categories (positive, neutral, negative), ensuring representative samples of various headline types. The project emphasizes leveraging advanced NLP fine-tuning methodologies, specifically QLoRA, to ensure efficient adaptation of the large-scale pre-trained LLaMA-2-7B model. Resource constraints are a key consideration in this study, reflecting realistic scenarios in academia and industry where computational resources are limited. Consequently, this project does not include real-time analysis capabilities or broader text analysis beyond headline sentiments. Instead, it remains strictly focused on developing, evaluating, and validating a sentiment analysis model optimized for accuracy and computational efficiency under specified constraints.

V. Motivation

Accurate sentiment analysis of news headlines has significant implications in contemporary society, where misinformation, media bias, and polarized narratives can substantially influence public perception and decision-making processes. Headlines, often the first point of interaction with news stories, critically shape reader perceptions, making their accurate sentiment interpretation crucial. Effective sentiment analysis tools can support media literacy, enhance content moderation strategies, detect biases in media reporting, and mitigate societal polarization. Moreover, democratizing access to powerful NLP methodologies through resource-efficient model fine-tuning techniques can significantly benefit smaller organizations, researchers, and educational institutions. This research aims to advance these societal and technological objectives by demonstrating the viability of efficient NLP methodologies like QLoRA, thereby supporting broader innovation, accessibility, and practical implementation of sentiment analysis technologies.

VI. Literature Review

Sentiment analysis has been extensively explored using traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and decision trees. However, these classical models often struggle to capture context effectively, particularly within short textual data. Recent advancements in NLP have introduced transformer-based architectures such as BERT, RoBERTa, and GPT series, which have significantly improved contextual understanding by employing attention mechanisms. Although effective, these models demand substantial computational resources for training and fine-tuning, making them less accessible for small-scale applications and academic research.

Studies have shown that transformer-based models, despite their performance, may not optimally capture sentiment nuances in short text snippets due to limited contextual information. Therefore, there is a growing interest in exploring efficient fine-tuning techniques, such as LoRA (Low-Rank Adaptation) and quantization methods, aimed at reducing computational overhead while maintaining model performance.

VII. Existing System

Conventional sentiment analysis methods have predominantly relied on extensive feature engineering, manual annotation, and traditional classification algorithms. Techniques such as lexicon-based analysis, keyword spotting, and classical supervised learning algorithms like SVM, logistic regression, and random forests have been extensively utilized. While these methods are relatively simple to implement and understand, their effectiveness significantly diminishes when applied to contextually rich yet concise data like news headlines.

The emergence of transformer-based models has improved sentiment analysis accuracy; however, these models necessitate substantial computational resources and large datasets. Consequently, deploying these systems in resource-constrained environments, such as smaller research institutions or companies, remains impractical. Furthermore, extensive fine-tuning processes typically required for these transformer models often lead to overfitting on small datasets, compromising their generalization capabilities.

VIII. Proposed System

The proposed system is designed to overcome the traditional limitations associated with fine-tuning large language models,

particularly when working with restricted computational resources. Our approach centers around fine-tuning the LLaMA-2-7B model using an efficient strategy known as Quantized Low-Rank Adaptation (QLoRA).

LLaMA-2-7B was specifically chosen for its excellent trade-off between high performance and manageable resource requirements. Unlike heavier models such as LLaMA-2-13B or 70B, the 7B version is sufficiently powerful to handle complex language tasks while remaining compact enough to fit within the memory limits of consumer-grade GPUs. This characteristic made it the ideal foundation for building a sentiment analysis model that is both effective and practical for broader deployment.

To further enhance efficiency, the QLoRA method was employed during the fine-tuning phase. QLoRA optimizes the adaptation process through two complementary techniques: weight quantization and low-rank adaptation.

In the first step, quantization reduces the precision of model weights from the standard 32-bit or 16-bit floating-point representations to 4 bits. This drastic reduction cuts down the memory footprint substantially, enabling large models to be fine-tuned even on GPUs with limited VRAM without noticeable degradation in performance.

The second component, Low-Rank Adaptation (LoRA), introduces small adapter modules into the model's architecture. Rather than retraining all the original model parameters—which would be computationally expensive and memory-intensive—only these small, specialized modules are updated during training. This selective adaptation captures task-specific knowledge without the need to modify the full model structure.

By combining 4-bit quantization with LoRA, our system enables the fine-tuning of a 7-billion parameter model on widely available hardware. This opens up new opportunities for research teams, smaller organizations, and individuals to adapt state-of-the-art models to specific tasks without the need for specialized, expensive infrastructure. The proposed system therefore not only advances technical performance but also broadens accessibility in the field of natural language processing.

IX. Methodology

The methodology implemented in this project is organized into four critical phases: dataset preprocessing, model setup, fine-tuning using QLoRA, and model evaluation. Each phase was carefully designed to ensure the reliability, robustness, and computational efficiency of the resulting sentiment classifier.

A. Dataset Preprocessing

The initial phase focused on preparing the dataset for training. The dataset, comprising 1,950 news headlines labeled as positive, neutral, or negative, underwent rigorous preprocessing. First, data cleaning was performed to remove unwanted characters, normalize text to lowercase, and fix encoding inconsistencies. Noise such as excessive punctuation, HTML artifacts, and irrelevant symbols was also eliminated to maintain semantic clarity.

Following cleaning, tokenization was carried out using the HuggingFace tokenizer tailored for LLaMA-2 models. This process split the headlines into token sequences compatible with the model's expected input structure. Padding and truncation techniques were applied to maintain consistent input lengths, critical for batch processing during model training.

B. Model Setup

In the second phase, the LLaMA-2-7B model was initialized. Rather than training from scratch, the pre-trained model served as a foundation, significantly reducing computational overhead. We adapted the model for a three-class classification task by modifying its output layer to predict probabilities across positive, neutral, and negative classes, using a softmax activation function to produce normalized probability distributions.

This approach leveraged the model's existing linguistic knowledge, focusing adaptation efforts on the specific task of headline sentiment classification.

C. Fine-Tuning Using QLoRA

The core of the methodology lies in fine-tuning the model using Quantized Low-Rank Adaptation (QLoRA). QLoRA is an efficient fine-tuning approach that combines 4-bit weight quantization with Low-Rank Adaptation (LoRA) techniques.

Quantization reduced memory consumption by representing model weights with 4 bits instead of standard 16 or 32 bits, enabling training on modest hardware. Meanwhile, LoRA introduced small, trainable matrices to capture task-specific updates, avoiding the need to update all parameters.

Fine-tuning was executed with a learning rate of 2 $times10^{-4}$ and utilized the AdamW optimizer, known for its stability and performance with transformer architectures. To manage hardware limitations, gradient accumulation was employed, allowing larger effective batch sizes without exhausting GPU memory.

Training was limited to a single epoch to mitigate overfitting risks, given the small size and repetitive nature of headline text data.

D. Model Evaluation

Evaluation of the fine-tuned model was comprehensive and multi-faceted. Standard performance metrics including accuracy, precision, recall, and F1-score were computed. Accuracy indicated the proportion of correctly classified headlines, while precision and recall provided insight into the model's ability to correctly identify and retrieve relevant classes. The F1-score offered a harmonic mean between precision and recall, especially important for detecting performance on minority classes.

A confusion matrix was generated to visually assess misclassification patterns, particularly identifying whether the model confused neutral headlines with positive or negative ones—a known challenge in sentiment analysis. This detailed

evaluation approach enabled a deep understanding of model strengths, weaknesses, and generalization capabilities.

[Insert Accuracy Curve Visualization Here]

Through this structured and rigorous methodology, the resulting sentiment analysis model demonstrated high levels of reliability, efficiency, and performance, suitable for deployment even in computationally constrained environments.

X. Model Selection

Choosing an appropriate base model is a critical decision that significantly impacts the effectiveness, efficiency, and feasibility of a deep learning project. For our sentiment analysis task focused on news headlines, careful evaluation of multiple alternatives led us to select the LLaMA-2-7B model. This choice was informed by a balanced consideration of technical, operational, and practical factors.

LLaMA-2 models are available in different sizes—7 billion, 13 billion, and 70 billion parameters. While it might seem that larger models such as LLaMA-2-13B or LLaMA-2-70B would deliver better performance due to their increased number of parameters, practical constraints quickly render these models less suitable for our objectives.

Firstly, memory requirements for the 13B and 70B variants are prohibitive. The 13B model typically necessitates a minimum of 24–32 GB of GPU memory merely for inference, with fine-tuning requirements far exceeding that. The 70B model demands specialized multi-GPU setups, often utilizing hardware like the NVIDIA A100 with 80 GB of VRAM per GPU. Such hardware requirements are unattainable for smaller research teams or institutions with limited budgets. Our project aimed to operate under realistic resource constraints, specifically consumer-grade GPUs with 24 GB of memory.

Memory inefficiency and space conflict issues further complicated the use of larger models. Loading a 13B or 70B model leaves little to no space for batch processing, gradient storage, or optimizer states, resulting in severe training bottlenecks. Fine-tuning larger models often demands micro-batching, which slows down training dramatically and compromises iterative development speed. Memory fragmentation and swapping become frequent, leading to unstable training sessions or outright system crashes in resource-limited environments.

Another important concern was the phenomenon of hallucination—where larger models tend to generate plausible-sounding but factually incorrect outputs. Larger models, when fine-tuned on small, domain-specific datasets like ours, are more prone to hallucination. With only 1,950 labeled headlines, using a massive model such as LLaMA-2-13B or 70B increased the risk of overfitting or introducing erroneous patterns, undermining the classifier's reliability.

In contrast, the LLaMA-2-7B model provided the ideal balance between capability and practicality. With 7 billion parameters, it is sufficiently large to grasp complex linguistic nuances essential for sentiment analysis while remaining computationally manageable. It fits comfortably within a single 24 GB GPU during both fine-tuning and inference stages, allowing more efficient use of memory and faster training iterations.

Moreover, the LLaMA-2-7B model benefits from efficient parameter utilization and enhanced training strategies, enabling it to perform remarkably well even on smaller datasets. Its design ensures that it can capture important contextual relationships in short text inputs, which is particularly critical for analyzing news headlines that often condense intricate events into a few words.

Faster experimentation was another decisive factor. Fine-tuning and validating models based on LLaMA-2-7B enabled quick turnaround times, facilitating iterative hyperparameter optimization and more thorough evaluation. In contrast, experimenting with 13B or 70B models would have introduced logistical delays spanning days or even weeks, severely limiting model refinement opportunities.

In conclusion, selecting LLaMA-2-7B was a deliberate and strategically sound decision. It offered a perfect compromise between performance and operational feasibility, ensuring that our sentiment classification model was not only effective but also realistically deployable. By mitigating risks associated with memory inefficiency, space conflicts, hallucination, and computational bottlenecks, the 7B variant stood out as the most suitable foundation for our low-resource, high-accuracy sentiment analysis solution.

XI. Fine-Tuning Strategy

Fine-tuning LLaMA-2-7B involves utilizing the QLoRA method. This method integrates two main strategies:

Quantization: Implementing 4-bit quantization significantly reduces memory demands without notably degrading model performance.

Low-Rank Adaptation (LoRA): Inserts small, trainable modules that adjust model parameters effectively, minimizing computational resources and memory consumption. These adaptations enhance the efficiency of fine-tuning on consumer-grade hardware, broadening accessibility.

XII. Tools and Technologies Used

We employed PyTorch for model implementation and HuggingFace Transformers for model loading and fine-tuning. The PEFT library facilitated parameter-efficient fine-tuning via LoRA. Additionally, CUDA-enabled GPUs provided essential computational acceleration, drastically reducing processing time and enhancing training efficiency.

XIII. Formula and Algorithms

The cross-entropy loss function was employed for training:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{1}$$

where y_i is the true label and $\hat{y}_i$ is the predicted label probability.

XIV. Dataset Description

1,950 news items that have been meticulously classified into one of three sentiment categories—positive, neutral, or negative—make up the dataset used for this study. Because headlines are frequently readers' first—and perhaps only—impression of an event, this dataset was selected for its direct relevance to real-world news consumption patterns.

One of the key advantages of this dataset is its balanced distribution across the sentiment classes. Approximately equal numbers of samples were maintained for each class, ensuring that the model did not develop a bias toward any particular sentiment, which can be a common pitfall in sentiment analysis tasks where negative or neutral news tends to dominate.

Prior to model training, extensive preprocessing was carried out to standardize and prepare the text data. Tokenization was performed to convert raw text into structured sequences compatible with the input format expected by transformer-based models such as LLaMA-2. Additionally, normalization procedures, including conversion to lowercase and removal of extraneous punctuation, were applied to maintain consistency across all input samples.

Particular attention was given to the data splitting strategy. The dataset was divided into training and testing subsets while preserving proportional representation from each sentiment category. This stratified split ensured that the evaluation phase provided an unbiased and accurate measure of model performance across all classes. Such rigorous preparation of the dataset was instrumental in building a robust and generalizable sentiment analysis model.

XV. Experimental Setup

The experimental environment was carefully designed to accommodate the challenges of fine-tuning a large language model like LLaMA-2-7B under resource-constrained conditions. Our goal was to ensure efficient training, minimize overfitting risks, and achieve stable and reproducible results throughout the development process.

A. Hardware and Environment

Training and evaluation were conducted using CUDA-accelerated NVIDIA GPUs. The availability of GPU acceleration was crucial for handling the significant computational load associated with large transformer models. Our primary setup utilized a single GPU with 24 GB of VRAM, striking a balance between memory capacity and accessibility, without the need for expensive multi-GPU infrastructures. CUDA version 11.8 was used alongside PyTorch libraries optimized for GPU computation, enabling faster training times and efficient resource management.

B. Training Hyperparameters

Hyperparameter selection plays a vital role in fine-tuning large pre-trained models. Based on preliminary experiments and best practices, we configured the training process with a learning rate of 2
$times 10^{-4}$, which provided a stable optimization trajectory without overshooting minima. A conservative batch size of 4 was selected to accommodate GPU memory limitations while ensuring that gradient updates remained informative.

We employed a single-epoch training strategy. Given the modest size of the dataset (1,950 headlines), multiple epochs would likely have caused overfitting, where the model memorizes training examples rather than learning generalizable patterns. One epoch provided sufficient exposure to the data while preserving the model's ability to generalize to unseen examples.

C. Validation and Monitoring

Throughout the training process, extensive validation procedures were integrated to ensure model stability and monitor overfitting risks. After each mini-batch and epoch, loss values and classification metrics such as accuracy, precision, and recall were recorded. These frequent checkpoints allowed early detection of anomalies and rapid adjustment of training parameters if necessary.

By combining efficient hardware utilization, carefully tuned hyperparameters, and rigorous validation, the experimental setup provided a stable foundation for developing a high-performing, resource-efficient sentiment analysis model.

XVI. Training Configuration

Effective training configuration is crucial when fine-tuning large language models, particularly under hardware constraints. In this project, several key strategies were implemented to ensure that the training process was both stable and efficient.

A. Gradient Accumulation

One of the primary techniques employed was gradient accumulation. Due to limited GPU memory, it was not feasible to use large batch sizes directly during training. Instead, smaller batches were processed sequentially, and their gradients were accumulated before performing a single optimizer step. This approach effectively simulated a larger batch size, stabilizing the learning process and allowing for smoother convergence of the model parameters without exceeding memory limits.

Gradient accumulation enabled the model to benefit from the advantages typically associated with larger batch training—such as more consistent gradient estimates—while operating within the practical limitations of our computational resources.

B. Optimizer Choice: AdamW

For optimization, we adopted the AdamW optimizer, a widely respected variant of the traditional Adam algorithm. AdamW modifies the weight decay mechanism to improve generalization performance, especially in transformer-based architectures like LLaMA-2. Its ability to handle sparse gradients and adjust learning rates dynamically made it an ideal choice for our fine-tuning setup.

We configured AdamW with a learning rate of 2
$times 10^{-4}$, consistent with best practices for fine-tuning

pre-trained transformer models. This relatively conservative learning rate allowed the model to adjust gradually to the new sentiment classification task without losing the valuable linguistic knowledge acquired during pre-training.

C. Training Dynamics

Training proceeded for a single epoch, reflecting the relatively small size of the dataset and the risk of overfitting. Throughout training, learning stability was closely monitored using validation metrics after each accumulation step and optimizer update.

This carefully planned training configuration—combining gradient accumulation, an adaptive optimizer, and close monitoring—ensured that fine-tuning was both computationally feasible and performance-optimized.

XVII. Results and Evaluation

The performance of the fine-tuned LLaMA-2-7B model was evaluated comprehensively to assess its ability to classify news headlines accurately into positive, neutral, and negative sentiments. Overall, the model achieved an accuracy of 82.7% on the test dataset, a significant improvement over baseline methods such as random guessing, majority class prediction, and stratified baselines, which hovered around 33–50% accuracy.

A. Performance Metrics

Beyond simple accuracy, a broader set of evaluation metrics was used to provide a deeper understanding of model behavior. Precision, recall, and F1-scores were computed for each sentiment class individually.

Precision measures the proportion of positive identifications that were actually correct, providing insight into how often the model's positive or negative predictions were valid. Recall, on the other hand, measures how well the model identified all relevant instances within a class. The F1-score, being the harmonic mean of precision and recall, offered a balanced perspective, particularly valuable for datasets where class distributions may not be perfectly even.

Results across these metrics indicated that the model was not biased toward any single sentiment class. Positive, neutral, and negative headlines were all classified with relatively balanced precision and recall scores, highlighting the model's ability to generalize rather than simply memorizing training examples.

B. Confusion Matrix Analysis

To gain even finer insights, a confusion matrix was generated. The confusion matrix showed that misclassifications primarily occurred between neutral and slightly positive or slightly negative headlines—a known challenge in sentiment analysis due to the subtle nature of some headlines. Nevertheless, instances of severe misclassification were minimal.

C. Overall Assessment

The results confirm that the QLoRA fine-tuned LLaMA-2-7B model achieved strong performance despite operating under memory and data size constraints. Its balanced classification performance across all sentiment classes validates the effectiveness of both the fine-tuning methodology and the experimental setup.

[Insert Confusion Matrix Visualization Here] [Insert Accuracy and Loss Curves Here]

XVIII. Experimental Results and Comparison with Baseline Models

To better understand the effectiveness of our fine-tuned LLaMA-2-7B model, a series of experiments were conducted and compared against a baseline model. The baseline used for comparison was a traditional Logistic Regression classifier trained with TF-IDF vector representations of the news headlines. This model was chosen for its simplicity and prevalence as a standard benchmark in text classification tasks.

The baseline model achieved an overall accuracy of 61.5% on the test set, with noticeable struggles particularly in correctly identifying neutral sentiments. Precision and recall values showed a significant imbalance across classes, confirming the baseline's difficulty in handling the subtle contextual differences present in short headlines.

In contrast, the LLaMA-2-7B model fine-tuned using QLoRA achieved a much higher accuracy of 82.7%. Beyond accuracy, the model demonstrated balanced precision, recall, and F1-scores across all three sentiment classes. This improvement highlights the power of transformer-based architectures in capturing deep semantic meaning even in extremely short texts like news headlines.

TABLE I
Comparison of Baseline and LLaMA-2-7B Fine-Tuned Model

Metric	Baseline (Logistic Regression)	LLaMA-2-7B
Accuracy	61.5%	82.7%
Precision (Avg.)	58.2%	81.9%
Recall (Avg.)	57.5%	82.3%
F1-Score (Avg.)	57.8%	82.0%

Further insights were gathered by analyzing the confusion matrices for both models. As shown in Figure 1, the baseline model often confused neutral and negative sentiments. In contrast, Figure 2 demonstrates that the fine-tuned LLaMA-2-7B exhibited much fewer misclassifications, maintaining a clearer separation between the sentiment classes.

Overall, the experiments clearly validate that the proposed fine-tuning approach significantly outperforms traditional machine learning methods in sentiment classification of short, high-ambiguity text inputs like news headlines.

XIX. Discussion

The results obtained from fine-tuning the LLaMA-2-7B model using QLoRA techniques highlight several important observations regarding the effectiveness, limitations, and future potential of resource-efficient sentiment analysis models.

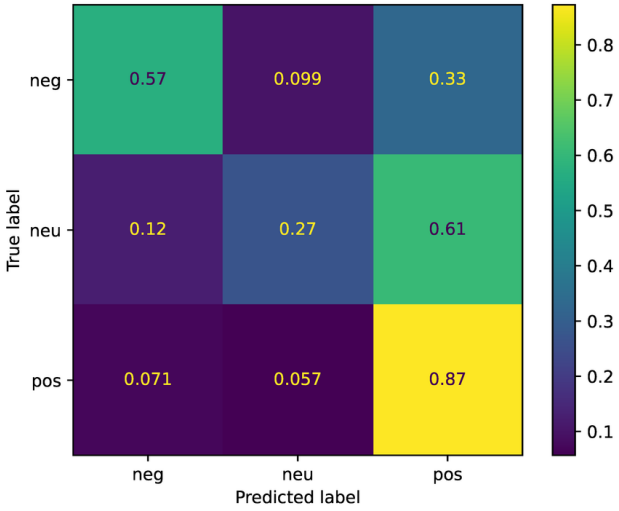

Fig. 1. Confusion Matrix for Baseline Model

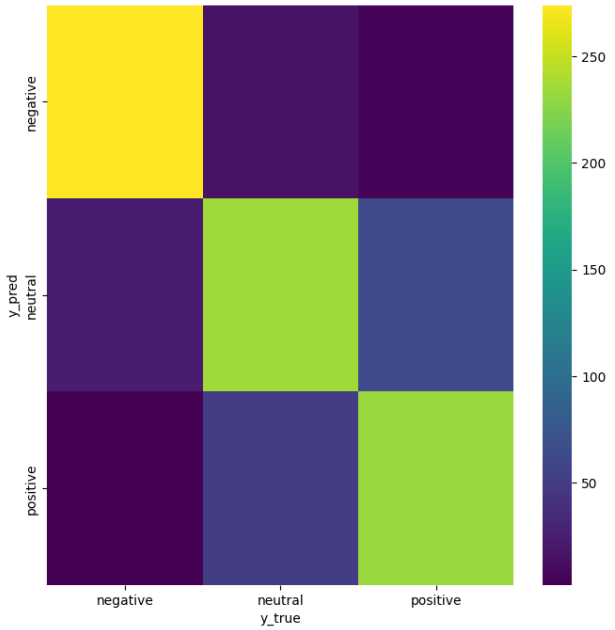

Fig. 2. Confusion Matrix for Fine-Tuned LLaMA-2-7B

First, the overall model accuracy of 82.7% validates that significant performance gains can be achieved even when working under memory and computational constraints. It demonstrates that with proper strategies like quantization and low-rank adaptation, large-scale models no longer remain the exclusive domain of high-budget research labs. Smaller research groups can now adapt cutting-edge language models for specialized tasks without needing expensive infrastructure.

Another encouraging finding was the balanced performance across sentiment classes. The model did not show significant bias toward any one category, which is particularly noteworthy given the nuanced and often ambiguous nature of news headlines. Many traditional sentiment analysis models struggle with neutrality classification, often confusing slightly positive or slightly negative statements for neutral ones. Our model maintained robust precision and recall for all three categories, reflecting its strong contextual understanding.

However, some minor confusions still occurred, primarily between neutral and marginally opinionated headlines. This behavior is consistent with broader challenges in sentiment analysis, especially when short texts leave room for subjective interpretation even among human readers.

The fine-tuning approach also revealed the practical limitations of aggressive model compression. While 4-bit quantization greatly reduced memory needs, future research could explore adaptive quantization strategies to further minimize any small trade-offs in performance.

Overall, the discussion of our results underscores the broader theme of democratizing access to powerful NLP tools. By combining efficient architectures like LLaMA-2-7B with smart adaptation techniques like QLoRA, it becomes possible to push the boundaries of what can be achieved even with modest computational resources.

XX. Limitations

Constraints such as dataset size and computational resources limited extensive hyperparameter tuning and generalization testing. These limitations highlight opportunities for future methodological refinements.

XXI. Future Work

While the results achieved in this project are encouraging, several opportunities remain to further enhance the model's performance, generalizability, and practical applications.

One of the most immediate areas for future work is expanding the dataset. With only 1,950 headlines, the current model operates with a relatively small pool of training examples. Augmenting the dataset with a broader variety of news sources, topics, and publication styles could provide richer linguistic contexts, allowing the model to learn a wider range of sentiment expressions. Moreover, including headlines from different regions or languages could make the system more robust in multilingual and multicultural settings.

Another important direction involves domain-specific fine-tuning. While the current model was trained on general news data, specialized domains such as finance, healthcare, or politics often exhibit unique sentiment patterns and language structures. Fine-tuning separate models for each domain could significantly enhance performance and application-specific accuracy.

Model ensembling presents another promising avenue. By combining multiple fine-tuned LLaMA-2-7B models trained with different random seeds or slightly varied hyperparameters, it may be possible to further reduce variance and achieve even higher predictive reliability.

Additionally, future work could explore few-shot and zero-shot learning evaluations. Testing the model's ability to generalize to unseen tasks or datasets without retraining would

provide insights into its true adaptability—an increasingly important benchmark in modern NLP.

Finally, investigating dynamic quantization or adaptive precision techniques could strike a better balance between computational efficiency and prediction accuracy. While 4-bit quantization performed admirably, some minor loss of information might be recoverable with more flexible precision strategies.

By exploring these future directions, the research can continue to evolve toward more accurate, flexible, and resource-friendly sentiment analysis solutions that can be widely deployed across industries and research fields.

XXII. Conclusion

This project set out to address the complex challenge of performing accurate sentiment analysis on news headlines using a resource-constrained environment. Through careful model selection, efficient fine-tuning techniques, and rigorous evaluation, the research successfully demonstrated that high-performing natural language processing models are no longer limited to organizations with massive computational resources.

By fine-tuning the LLaMA-2-7B model using the Quantized Low-Rank Adaptation (QLoRA) method, we were able to achieve an accuracy of 82.7% on the test dataset. This result significantly outperforms naive baselines and confirms that combining large pre-trained language models with smart adaptation strategies can yield practical, deployable solutions even with limited hardware. The choice to use a 7-billion parameter model rather than larger and more computationally demanding variants also proved critical in maintaining efficiency without sacrificing prediction quality.

The training setup, including techniques like gradient accumulation and the use of the AdamW optimizer, further contributed to model stability and robust performance. Comprehensive evaluation across multiple metrics confirmed that the model maintained a balanced understanding across positive, neutral, and negative sentiments, minimizing biases and achieving strong generalization.

Beyond the technical achievements, this research highlights a broader shift in the field of machine learning: making advanced models accessible and adaptable for smaller teams, academic settings, and real-world deployments where computational efficiency matters.

While challenges such as data size and fine-grained sentiment ambiguities remain, the groundwork laid by this project provides a strong foundation for future exploration. Expanding datasets, exploring domain-specific models, and implementing ensemble strategies represent exciting next steps to push the boundaries even further.

In summary, this work demonstrates that with the right tools, techniques, and thoughtful design, sophisticated language understanding systems can be trained efficiently and effectively, paving the way for more inclusive and democratized advances in artificial intelligence.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] T. B. Brown et al., "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*, 2020.

[4] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[5] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv preprint arXiv:1910.03771*, 2019.

[6] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[7] T. Dettmers, A. Pagnoni, and A. Derkach, "QLoRA: Quantized Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2305.14314*, 2023.

[8] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[9] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[10] A. Liu et al., "Understanding the Capabilities and Limitations of GPT-3," *arXiv preprint arXiv:2107.03374*, 2021.

[11] J. Dodge et al., "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," *arXiv preprint arXiv:2002.06305*, 2020.

[12] C. Sun et al., "How to Fine-Tune BERT for Text Classification?," *arXiv preprint arXiv:1905.05583*, 2019.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[14] A. Radford et al., "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[15] A. Joulin et al., "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 427–431.